Winter 12-10-2022

Developing App from User Feedback using Deep Learning

Chengfei Wang
Auburn University, chengfeiwang.aka.chance@gmail.com

Xiao Qin
Auburn University, xqin@auburn.edu

Ashish Gupta
Auburn University, azg0074@auburn.edu

Developing App from User Feedback using Deep Learning

Abstract

The mobile app market, a multi-billion-dollar industry, is highly competitive. Online consumer feedback can provide good insights into product strengths and weaknesses. In this study, we utilize deep unsupervised learning models to build a framework for harnessing potential customer feedback. The first part of the framework uses Bidirectional Encoder Representations from Transformers (BERT)-based topic modelling approach to identify topics and key themes that emerge from user reviews of mobile apps belonging to the health and fitness genre. The second part, sentiment analytics, integrates the accompanying ratings to reveal the market acceptance of various aspects of the product design. The findings provide strong guidance for improving the design and development of such apps. The study has important implications for creating an AI-driven app design framework.

Keywords

BERT, Mobile App Reviews, Topic Modelling.

Introduction

The mobile phone application market is highly competitive and has now grown into a multi-billion-dollar industry. The popularity of these apps follows a typical power law distribution: more than 95% apps are downloaded by fewer than 1,000 devices, while few apps receive over a million downloads (Liu et al. 2017). A recent study (Rob van der Meulen 2014) by Gartner supports this trend as it predicts that less than 0.01% of consumer mobile apps would be considered a financial success in the future. Therefore, it is important for app developers to understand the characteristics of a successful mobile apps which could draw consumer's attention. The design of a successful software product is a core research topic in quality assurance. Various metrics have already been proposed to promote the software quality (Gaffney Jr 1981). Although inspiring, those metrics don't provide specific design guidelines.

Consumer reviews posted on app distribution sites such as google store could provide good insights into various strengths and weakness of apps. Several reasons may contribute the wide adoption and usage of a mobile app or its failure. Unlike other digital products such as music and movies that are often sold as finished products, mobile apps as software products offer developers an opportunity to integrate customer feedback from similar apps into various design stages for improving future app functionalities. Recent studies have reported that later app releases, which typically have improved design features and functionalities, tend to have greater influence on an app's success in terms of higher sales performance than early releases (Lee and Raghu 2014). Improvement in app features could benefit from deeper understanding of specific customer needs. Analysis of these mobile app reviews provide such understanding.

Although mobile app reviews are appropriate subjects for text mining due to their relatively short lengths, limited scopes, large quantity and tremendous product insights contained, however, analysis of them is still a challenging task. Information such as bug report, overall user experience or new feature requests are usually mixed up in user reviews (Maalej and Nabil 2015). Labels of the data are usually not available. A model trained from one labelled dataset cannot be applied on other datasets due to the different genre characteristics. It is also difficulty to process and analysis comments manually due to the large scale. An unsupervised framework automatically extracts detailed semantic information should be developed to overcome the problem mentioned above.

To extract useful information from reviews, most application tools such as competitive analysis, summarizing, classification of mobile app reviews in the literature are developed based on part-of-speech rule-based (Johann et al.) techniques or topic modelling (Blei et al.). However, the semantic meaning of

extracted words that depends on context was ignored. The recent rising of deep unsupervised learning model lifts the nature language processing to a new level. Pre-trained model such as BERT (Devlin et al.) are context-aware and can convert semantically similar sentences to well-clustered vectors, which allows us to build a system that sensitive and better understand user feedback.

This study has three major objectives. One is to review prior research related to the strategies used to build automatic analysis system. The second objective is to compare and contrast popular and unpopular features among certain genres of apps and identify the determinants of successful mobile app design. We use the health and fitness category in Google Play Store as our example, but the framework can be generalized to any category of mobile apps. The third objective is to extend the analysis results to new app design guidelines that help developer avoid risks and prioritize the development of the most important app features. Those guidelines include recent strategies such as influencer marketing and gamification. Currently, few studies in the literacy has extended focus to complete guidelines of a success mobile app and beyond the scope of software quality.

Literature Review

Texting mining has been extensively used for developing insights from customer reviews. While a plethora of research on customer reviews focuses on traditional text mining approaches, more recent research is utilizing deep learning-based text analytic models. We performed extensive review of extant literature and noticed that a majority of studies could be classified into one or more of the five groupings based on their study objectives as summarized in Table 1. These are (1) feature extraction, (2) classification, (3) clustering, (4) ranking and (5) summarization. We will, first, briefly discuss important research done in each of these five areas and subsequently describe deep learning-based approaches as applied to understand customer reviews.

A majority of research on feature extraction focuses on identifying one or more important features, such as product-based features (for e.g., product type) (Liu et al. 2021), lexical features (for e.g., part of speech tags) (Malik et al.), linguistic features (for e.g., review length) and others. Lexical-based feature extraction methods follow lexical rule library or keyword dictionary to extract certain words from review text as features. They can also be used to score the sentiments of the reviews, such as SentiStrength (Shah et al.). Statistics-based feature extraction methods treat the text as a bag of words (BOW). The importance of words are measured by term frequency methods (Goldberg and Abrahams) and the relevance between words are measured by item-set frequency methods (Liu et al.). In practice, many take both statistics-based and lexical-based approaches. (Chen et al.).

Multi-classification divides reviews into categories such as feature request, bug report, feature evaluation, or others (Gu and Kim 2015; Maalej and Nabil 2015; Shah et al. 2019; Singh and Tucker 2017) as the initial step to understand customer intentions. Binary classification filters out reviews irrelevant to the goal of task, such as whether the review is informative (Chen et al. ; Zheng et al.). Keyword dictionary-based classifier can be built by statistical methods such as Fisher test (Zhu et al.). Manually crafted scores, such as feature performance score (Dalpiaz and Parente) are used as criteria for classification. Machine learning models, such as Naive Bayes, decision tree, K-nearest neighbour (KNN), and support vector machine (SVM) are used by many researchers (Chen et al. ; Singh and Tucker). Some researchers argued that multinomial logistic regression (Max Entropy) has the best performance on text classification (Gu and Kim ; Maalej and Nabil ; Shah et al.). Ensemble learning further improves the performance (Liu et al. ; Zheng et al.) by combining multiple classifiers.

Clustering plays a vital role on discovering patterns in a large number of reviews by grouping similar ones together. For example, researcher can convert reviews text into document vectors by TF-IDF then apply k-means method (Flory et al.). However, a document text may associate with multiple topics, which may require probabilistic graph methods. The topic modelling algorithms Latent Dirichlet allocation (LDA) is used by many (Chen et al. ; Fu et al. ; Guzman and Maalej ; Verkijika and Neneh ; Zhu et al.) and has best performance in terms of F-measure (Chen et al.). For short text, simple item-set mining methods are efficient (Gu and Kim ; Guzman and Maalej ; Shah et al.). Sets of terms with high support counts are grouped according to a synonym dictionary such as WordNet.

Ranking prioritize the results by relevance and importance. The measurement of relevance and importance could be the volume of reviews in each group (Dalpiaz and Parente ; Li et al.) or other engineered scores.

Ranking scores can be derived from the linguistic statistics of the text which measures similarity such as support count and the number of shared features (Shah et al.), or from the ratings by the reviewers since lower average ratings usually means urgent issue (Chen et al. ; Chen et al.), or from the sentiment analysis such as Average Sentiment Score (Liu et al.).

No.	Year	Author	feature	classification	clustering	ranking	summerization	inference	Key Findings
1	2022	Goldberg et al.	√	√		√	√		Smoke words extracted from reviews of feature request, irritator, compliment categories are useful
2	2021	Verkijika et al.			√		√		Ease of use, usefulness, convenience are positive themes, customer support, recieved cost, lack of trust are negative themes
3	2021	Ha et al.	√				√		Form morphological matrix from extracted keywords provides new idea of innovation
4	2020	Liu et al.	√	√			√	√	(1) consumers are not sensitive to the price of elderly phones, (2) but sensitive to the price of other smartphones, (3) wide and thin phones are more competitive in size
5	2020	Zheng et al.		√	√		√		(1) use three filters, namely the sentiment filter, the component-symptom filter and the similarity filter, to select informative threads (2) identifies the threads related to product defects and provides detailed defect information including defect types, defective components and defect symptoms
6	2019	Liu et al.	√	√		√	√		(1) seeding words can be expanded into domain-specific sentiment lexicon (2) identify comparative text and competitive product from forum, and aspect comparison information from another pre-categorized source
7	2019	Chen et al.	√	√	√	√	√		Suggestions ranking high by both count & rating have a higher probability of improving the upgrade
8	2019	Dalpiaz et al.	√	√			√		High review volume may reduce false positive case of feature extraction, human analytical skills
9	2019	Shah et al.	√	√			√		Categorizes review sentences into feature evaluation, bug report and feature request
10	2018	Malik et al.	√		√	√			Similar app represented in similar feature tree and be compared
11	2018	Liu et al.	√	√					Essemble learning of bagging can improves performance of identify product complain thread
12	2018	Zhu et al.	√	√	√		√		5 students evaluation show the designed system very concise and usefull to retival sentences
13	2018	Marcacini et al.	√						(1) providing a unified representation of feature spaces between different domains through heterogeneous transductive networks (2) using a cross-domain transfer learning process to propagate label
14	2018	Shah et al.	√	√					Classified into Praise, Feature Evaluation, Bug Report, Fearture request, other. The simple CNN model has comparative result to Max Entropy but much slower
15	2017	Singh et al.	√	√				√	Devideded reviews into behavior, form, function, service. for android phones, fucntion and form are positively related to ratings and behavior and services are negatively, decision tree J48 have good performance
16	2017	Johann et al.	√	√					Unfiltered user reviews reached average precision of 24% a recall of 71%. Features extracted from user reviews are still noisy but catch majority features discussed by the users.
17	2017	Di Sorbo et al.		√		√			Summarize app reviews and generate an interactive, structured and condensed list of recommended software changes by classification of intention and topic rank
18	2017	Li et al.	√	√				√	A deep learning-based approach for understanding and predicting users' rating behaviors unifying aspect ratings and review contents show good performance in rating prediction
19	2016	Flory et al.	√		√	√			Helpfulness is defined as relavance to customer search
20	2016	Shah et al.	√		√	√	√		Extract feature and sentiments from user reviews
21	2015	Maalej et al.	√	√					Naïve Bayes better than decision tree and Max Entropy and keywords baseline method, sentiment score is important, 4 binary classification better than 1 multi-classification
22	2015	Gu et al.	√	√	√		√		SUR-Miner provides reliable results on review classification, aspect-opinion extraction, and sentiment analysis, with each average F1-scores of 0.75, 0.85 and 0.80. SUR-Miner more focus than AR-Miner
23	2014	Li et al.	√	√		√		√	Combine all the feature generate the best product portfolios support by amazon best seller result
24	2014	Chen et al.		√	√	√	√		Topic modeling by LDA shows better performance than ASUM
25	2014	Guzman et al.	√		√		√		The extracted features were coherent and relevant to requirements
26	2013	Zheng et al.	√	√	√		√		(1) the social features of reviewers improve classification results (2) classification affected by product type due to the different purchase habits of consumers (3) reviews are contingent on the inherent nature of products, such as search goods or experience goods, digital products or physical products
27	2013	Fu et al.	√		√		√		The top-3 complaints around the same issues: content attractiveness, stability, and cost

Table 1. Literature Review

Summarization presents the final results in either text or numerical form, with additional visualization and inferences drawn from the data. Words contain rich semantic information, therefore, is the most common form (Fu et al. ; Goldberg and Abrahams ; Ha and Geum ; Verkijika and Neneh). Predefined templates turn short terms into well-organized sentences (Chen et al.). Tools from management science such as strengths,

weaknesses, opportunities, and threats matrix (SWOT) are very useful for competitive analysis (Dalpiaz and Parente) and innovation idea generation. Results may be reported in numerical values such as binary labels of useful reviews or not (Zheng et al.), or in text value pairs, such as the text of extracted aspects and values of sentiments scores (Guzman and Maalej). The numerical results can be visualized in bar-chart (Shah et al.), sized-dot plot (Shah et al.) or other forms. Linear regression (Li et al. ; Liu et al. ; Singh and Tucker), quantile regression, and mutual information (Liu et al.) are methods used to study the effect and contribution of each factor in the reviews.

The recent progress of deep learning improves the performance by larger and more complicated models. A deep neural network of attention mechanism was proposed to understand user's ratings shows good performance in rating prediction (Li et al.). Deeply Moving is another popular deep learning model for sentiment analysis (Gu and Kim). However, there is a paucity of literature on deep learning-based frameworks applied to understand customer reviews comprehensively. The traditional approaches of analysis on customer reviews have several limits such as tedious pre-processing, subjective feature engineering, and context unawareness. Deep learning models, especially transformer-based unsupervised deep learning models overcome those limits due to its strong ability of "understanding". Those models are trained without labels (unsupervised) under the setting of Language Mask Model, under which the models predict randomly masked words according to the context. The architecture of the neural network Transformer can draw attention to related surrounding words for prediction (Vaswani et al. 2017). Therefore, Transformer-based embedding neural network such as BERT are context-aware. This key advantage improves the overall understanding of the text and overcomes the issue of context understanding of traditional approaches. It greatly simplifies the workflow since it allows the model work directly on noisy raw text instead of pre-process and manual feature engineering.

Research method

The whole workflow of our framework and processed examples are illustrated in Figure 1. The framework used the transformer (BERT)-based methodology (Grootendorst 2022) to perform topic modelling and discover underlying topics/themes from collected user reviews, which includes major steps of data collection, pre-processing, embedding (vectorization), clustering, summarization (frequency mining and theme deduction) and sentiment analysis (tripartite graph).

Data Collection

We developed custom web page crawlers in Python for collecting user review data from the Google app store (Figure 1 crawling). Our data acquisition efforts were restricted to apps belonging to the health and fitness category. We further divided health and fitness category into 5 sub-categories. For each sub-category, we used 2 or 3 popular apps as search seeds to explore others. For each app, we collected the 100 most relevant comments. The app user reviews were collected over a wide range of time frames starting from each app's launch. The names of seed apps, number of apps and number of reviews of each sub-category are summarized in Table 9.

We utilized two different crawlers to scrape the app reviews efficiently. The first crawler used Selenium library to control the web browser (Chrome). The crawler extracted the information rendered by JavaScript in browser and stored the app list in a JSON file. The second crawler used an open-source python library (google play scraper) to connect with Google Play API directly and downloaded the review data of apps listed by the first crawler. We also used a paid proxy service (MeshProxy) to increase the concurrency of the crawling process. The two-crawler approach provides a key advantage in terms of ease-of-use and data collection efficiency.

Data Pre-processing

The performance of the deep unsupervised learning model we used (BERT) is not affected by the NLP pre-processing, however, the frequency mining of topic keywords requires the pre-processing. We applied several NLP pre-processing steps provided by python library NLTK to the raw text (Figure 1 pre-process). We first tokenized each review text and converted all into lower cases. Second, we eliminated the tokens in the stop word list from the NLTK library. Stop words are ubiquitous words in all documents without any distinguishing feature and semantic meaning, such as that, the, who, etc. In the third step, we used the

WordNet Lemmatizer of the NLTK to transform words into their basic form to merge the different variations of the same word. Finally, we filtered out unqualified reviews which are less than 2 words or non-English. Since the reviews collected from Google Play are already high quality and mostly English, simple filtering rule, which is more than the half letters must be English letters, is enough. Fig 1 illustrated examples of review before and after a series of pre-processing.

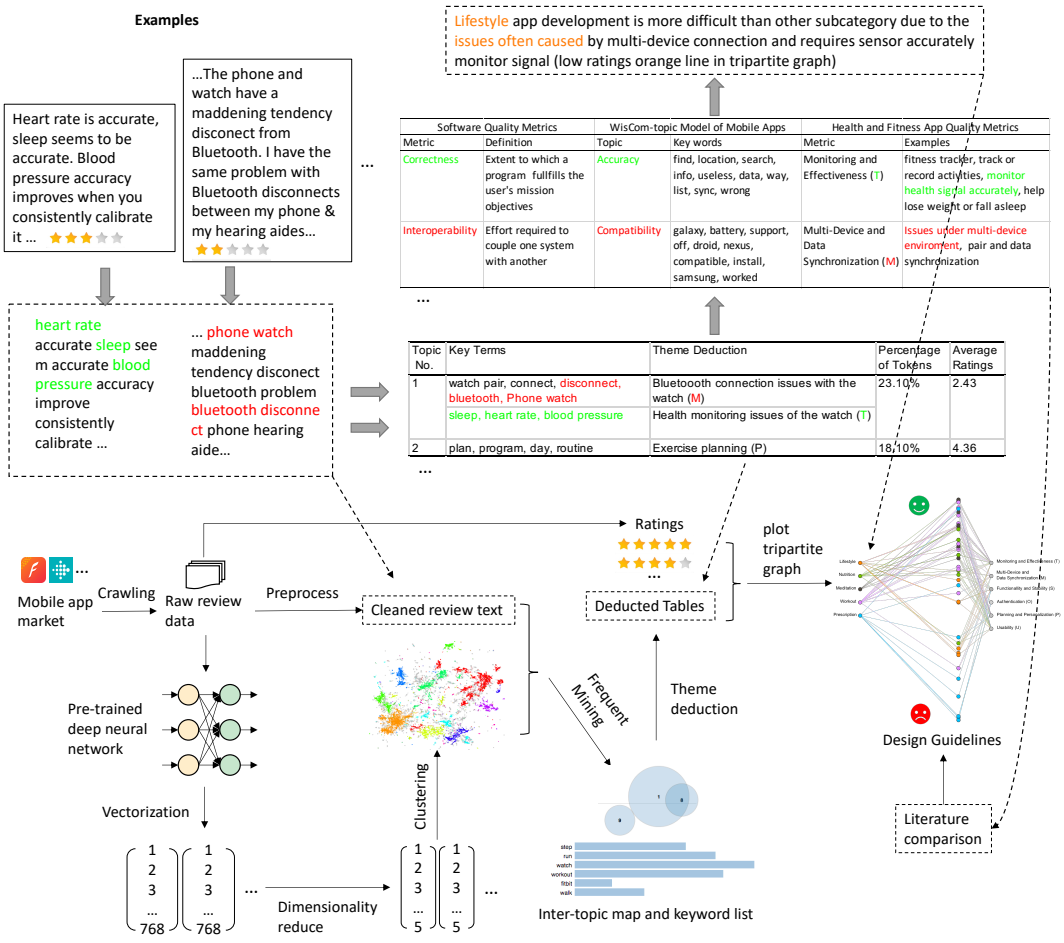

Figure 1. The Framework

Embeddings

To perform the embedding step, we use a pre-trained Sentence-BERT (SBERT) (Reimers and Gurevych 2019) model to convert review text into vector representations (Figure 1 vectorization). Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al. 2018) is an unsupervised deep learning framework. This contextual-aware transformer-based framework overcomes the semantical indistinguishability of words in different context by traditional embedding techniques that based-on simple bag-of-word features, and therefore shows large tolerance on processing noisy raw text and great performance in representing word semantics as vectors. SBERT is a variation of BERT and regarded as the state-of-the-art framework specialized on sentence embedding (Thakur et al. 2021). To our knowledge, no prior studies have used approach to study review texts of health and fitness mobile apps. Specifically, we use all-MiniLM-L6-v2 as our embedding model, a pre-trained model of SBERT on English corpus by Devlin et al. It maps each document into a 386-dimension semantically comparable dense vectors. We assume the vector representations of the semantically similar review text on the same topic are close to each other in

the vector space. The state-of-the-art pre-trained embedding model servers as the upstream for the downstream tasks of clustering, and the quality of topic modelling continuously grow as new state-of-the-art pre-trained models are developed.

Clustering

The high dimensionality of data is a major challenge of clustering. We applied the UMAP to reduce the dimensions of the learned representations from previous step (Figure 1 dimensionality reduce). As data increases in dimensionality, the difference between the distance to near and far data points decrease (Aggarwal et al. 2001; Beyer et al. 1999), therefore, the spatial locality becomes ill-defined. UMAP is a state-of-the-art method that preservers well both local and global features of high-dimensional data projected into lower dimensions (McInnes et al. 2018). We tune the hyper-parameters n = 15 (number of neighbours) and d = 5 (target embedding dimension) recommended by (Grootendorst 2022).

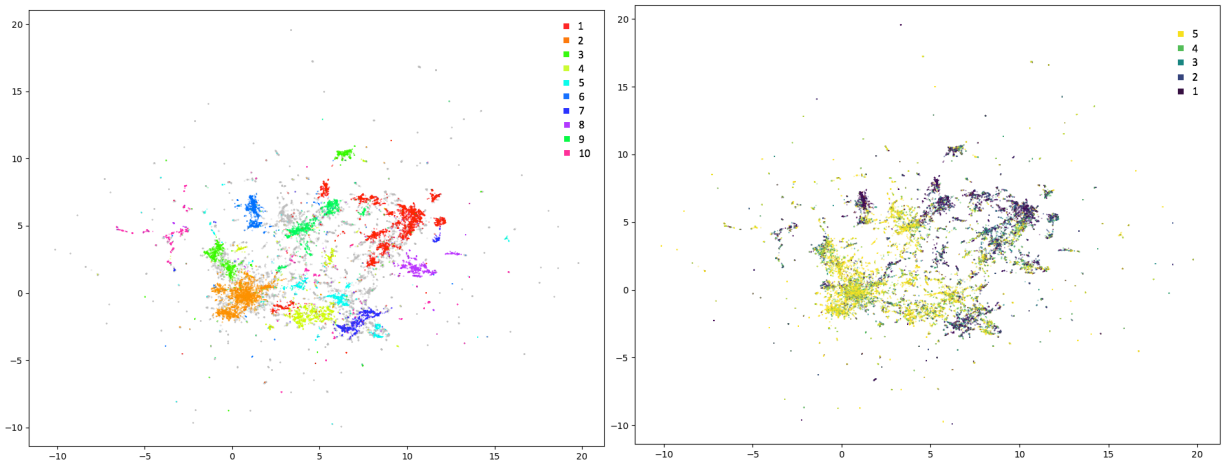

Figure 2. Clustering by (a) Topic and (b) Average Ratings

After dimensionality reducing, we use HDBSCAN algorithm (McInnes et al. 2017) to clustered embeddings, as suggested by (Allaoui et al. 2020). HDBSCAN is a single linkage hierarchical clustering algorithm on the transformed space adapted to the data point density using soft-clustering approach. It allows noise to be modelled as outliers and prevents multi-topic documents to be assigned to any single cluster. The minimal size of clusters is a hyper-parameter without any agreed formula to determine its optimal value. Therefore, we both qualitatively checked whether the selected value generates a meaningful set of topics and then quantitatively measured the distances among topics enough to separating different topics, as suggested by (Mortenson and Vidgen 2016). We used a web-based visualization package, PyLDAvis (Mabey 2018; Sievert and Shirley 2014) to check both from the bigram list and the inter-topic distance map (Figure 4). After several running, we choose the optimal values leading to minimal overlapping and meaningful topics for each subcategory of review texts, listed in Table 9 Column 5. Figure 2 shows an example of clustering result from lifestyle apps plotted in 2D space.

Summarizing

We summarized a topic with a list of words that selected from the collection of reviews that assigned to that topic. The words are either monograms or bigrams. Wallach pointed out that bigrams, such as 'New York', are better analysis units for topic modelling than single word tokens such as 'New' and 'York' since bigrams are able to better maintain the meaning and balance the dimensionality of the vocabulary constructed (Wallach 2006). We used multiple frequency-based techniques to mine semantically meaningful words. We first used c-TF-IDF, which is a modified TF-IDF approach formulated as following:

$$W_{t,c} = f_{t,c} \cdot \log\left(1 + \frac{A}{f_t}\right) \quad (1)$$

It is the product of two major parts, term frequency and inverse cluster frequency. We treat all documents in a cluster as one large concatenate document, therefore, the term frequency f is defined as the frequency

of term t in a cluster c. Similarly, the inverse cluster frequency is defined as the logarithm of the average number of words per cluster A divided by the frequency of term t across all the clusters. It adds 1 to the division within the logarithm to guarantee only output positive value.

We also used the *relevance* suggested by (Mabey 2018; Sievert and Shirley 2014) to measure the importance of words, which is defined as following:

$$r(t, c \mid \lambda) = \lambda \log\left(\phi_{t,c}\right) + (1 - \lambda) \log\left(\frac{\phi_{t,c}}{\phi_t}\right) \quad (2)$$

The $\phi_{t,c}$ denotes the probability of term t appears in the topic (cluster) c, while the ϕ_t denotes the marginal probability of term t in the corpus. The relevance calculated as the weighted sum of two logarithms of the topic-specific probability $\phi_{t,c}$ and the lift. The lift is a ratio of a term's probability within a topic (cluster) $\phi_{t,c}$ to its marginal probability ϕ_t across the corpus. Although the lift is useful to find important topic-specific words, it is also very sensitive to the rare terms. Combine both two logarithms avoids noisy results. Since the clustering result is available from previous steps, the probability ϕ can be simply approximated by frequency of terms.

The pyLDAvis visualization tool allows user to modify the parameters λ interactively, which controls weights of two logarithms in relevance. We checked the keyword list on right panel and the inter-topic distance map on the left panel (Figure 4). The size of a circle suggests the prevalence of that topic in the corpus by number of word tokens, and the distance between two circles reflects the similarities of the topics. For example, the biggest circle in Figure 4 represent Topic 1 is the most prevalent topic in lifestyle app sub-category, the circle of Topic 1 in Figure 4 (multiple device connection issues) is more similar to Topic 8 (data synchronization issues) than to Topic 9 (app crash).

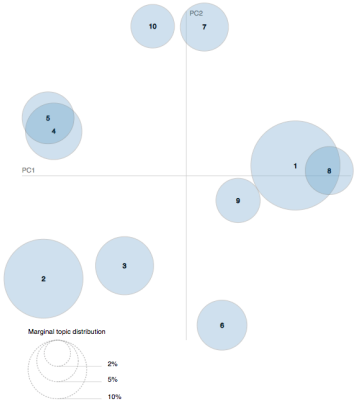

Figure 4. Inter-topic Distance Map

We collected and merged all the topic terms under three parameter values (λ=0.3, 0.6, 0.9). We manually selected one term for every synonym sets. Two researchers then independently went through topic terms to understand these latent topics and deduce themes within each topic. The process was repeated for all 5 subcategories. The themes generated from this process were then summarized in Table 2, 3, 4, 5, 6 and reported in the Results section.

Sentiment analysis

We combined the extracted topics and deducted themes with additional sentiment information, which is the rating scores comes along with each comment in scale of 5 stars. Many sentiment researches use user ratings as ground truth for sentiment prediction task (Fang and Zhan 2015), therefore, we use the ratings directly as measures of sentiment. The sentiment information provides better understanding of emotion embedded in the review text. Figure 2(b) shows vector representation of each review from lifestyle apps with each colour stands for different rating scores and Figure 2 with each colour for different clusters. We can differentiate opinions in similar topics by comparison. Topic 4 and 7 are close to each other in Figure 2, Topic 10 and Topic 7 are close in Figure 4. However, from the ratings we can easily tell Topic 7 complaints the frequent technical issue during the use while Topic 4 and 10 are praising the usefulness of it.

Experimental Results

We extracted 55 topics from the 5 sub-categories of the health and fitness app reviews from Google Play market. Table 2-6 show the theme deduction result of each sub-category. Since apps may be multifunctional and belong to multiple sub-categories, the categorization of the reviews is non-exclusive. We also assigned a tag for each theme and will discuss similar themes as groups in the next section.

Topic No.	Key Terms	Theme Deduction	Percentage of Tokens	Average Ratings
1	watch pair, connect, disconnect, bluetooth, Phone watch	Bluetoooth connection issues with the watch (M)	23.10%	2.43
	sleep, heart rate, blood pressure	Health monitoring issues of the watch (T)		
	notification, Update	Version stability (S)		
	watch face	Watch face customization issues (A)		
2	home workout, equipment, muscle group, great/good/best workout, full body, stretch, personal trainer, variety exercise, options	Guide to exercises (I)	18.10%	4.36
	plan, program, day, routine	Exercise plan (P)		
3	food item, nutrient, keto, carbs, gram, body fat, belly fat, sodium, eat, BMI, meal, recipe	Food and nutrition guide (I)	9.80%	3.68
	daily calorie, calorie macros, food track, kg/pound, measurement, lose weight	Weight monitoring and calorie counting (T)		
	barcode scanner	Convenient scanner (U)		
	scale, connect scale	Connection issues with the scale (M)		
4	runner, fasting, stats run, pace control, guided run, marathon, 5k	Guide for runner (I)	9.30%	4.00
	track pace	Track user's pace (T)		
	treadmill	Connection issues with the treadmill (M)		
	race	Exercise encouragement (E)		
5	find trails,	Information of trails (I)	7.80%	4.34
	live logging, calorie burn, step, walk, hike	Walk tracking (T)		
	daily walk	Walk planing (P)		
	motivated, goal, group ride	Exercise encouragement (E)		
	bike computer	Connection to the bike computer (M)		
	zwift, best cycling, easy	Cycling userability (U)		
6	cancel membership, collect coin, unsubscribe, auto renewal, charge free, redeem, try cancel, bank account, impossible cancel, refund policy, charge account, force pay, refund money, service use, premium package, free trial, customer service, purchase, email	Price and Customer Service (C)	7.20%	2.40
	full screen, click ad	Distruptive advertisement (U)		
7	mileage counter, distance, gps track, accurate	Location and distance tracking (T)	6.70%	2.50
	gps work, gps signal, signal lose, gps stop, update weather, weather location	Positioning stability (S)		
8	google fit, sync fitbit, mi fit, sync google, fit data, fit samsung, jyoupro, launcher, band	Google fit synchronization issues (M)	6.70%	3.20
	find fitbit, lose fitbit,	Locating lost device (U)		
	launcher	Launching issue (S)		
9	crash immediately, always crash, download open, feature compare, immediately open, say pending, start freeze, amount people, bug android, install reinstall, open already, won't open, load, try open	Application crash (S)	5.70%	3.21
10	pedometer, step counter, track step, sensitivity, accurate, simple/easy/basic pedometer	Pedometer tracking accuracy (T)	5.50%	3.54
	best/great pedometer	Pedometer usability (U)		

Table 2. Results of Lifestyle Apps

Results of Lifestyle App Reviews

The lifestyle app subcategory covers mostly tracking apps on all aspects of a healthy lifestyle. The size of 10 emerged topics from lifestyle app reviews, measured as the percent of tokens, ranges from 5.50% to 32.10% percent. The first topic of lifestyle app review is about the smart watch. It is the largest topic consisting of 4 themes that focus on the issues of connections between phones and smart watches, health signal monitoring issues, instability after software update and customization of watch faces. The second emerged topic consists of two themes focusing on guide of home workout and exercise planning. The third emerged topic mostly revolves around themes about healthy eating, such as food and nutrition guide, weight monitoring and calorie counting. Other themes are issues of the connection between smart scale and app and positive feedback on the convenient feature of barcode scanner (record food and provide related information by scanning). The fourth topic is about running exercises. The themes of fourth topic are the guide for runners, tracking user's running pace, encourage user by virtual racing and issues of connection to the treadmill. The fifth topic is about walk and cycling. The themes of the fifth topic are information of trails, tracking daily walk, planning, useability of cycling app and encouragements such as goal setting, group riding. Another theme is the issue of connection to the bike computer. The sixth topic focuses on charge disputes and disruptive advertisement experience. The seventh topic focuses on positioning issues, including themes of inaccuracy tracking and unstable service. The eighth topic focuses on the issues of Google fit data

synchronization, app launching issues and useful function of locating lost devices. The tenth topic focuses on pedometers, including themes of tracking their accuracy and usability.

The above analysis summarized in Table 2 provides specific guidelines for developing of certain type of app. For example, a success running exercise app should at least consider factors such as providing training guide, pace tracking, encouragement by holding virtual race and connection to treadmills. At the same time, we noted that the multi-device environment is the major challenge for lifestyle app developing. The quality of connection is a significant issue, including connecting to smart watches, smart scales, treadmills and bike computers. We also noted that a healthy lifestyle app may covers broad aspects of life, such as sleep, eat and workout, which will be discussed more in the following

Topic No.	Key Terms	Theme Deduction	Percentage of Tokens	Average Ratings
1	workout apps, gym, workout plan/program, equipment, posture, resistance band, stretch, muscle, finess, workout without	Guide to exercise muscle (I)	24.40%	4.4
	custom workout, daily yoga, routine, plan	Exercise planning (P)		
	deep sleep, workout tracker, rem sleep	Sleep monitoring (T)		
2	reset password, fitbit, sync samsung, fitbit versa, update, email, login, closing, fit sync, won't sync, account, forget password, crash, log, connect internet	Login and synchronization issues (S)	18.50%	3.62
	much ad, video ad	Disruptive advertisement (U)		
	edit food	Inconvenient editing (P)		
	track	Food tracker (T)		
3	cancal, subscription, charge, pay, refund, free trial, money, payment, customer service, premium, scam, try cancel, unsubscribe, paypal, credit card, bank account, google pay, try contact	Price and Customer Service (C)	9.90%	1.87
4	calorie counter, food, track calorie, eat, calorie intake, calorie burn, weight, calculate calorie	Food tracker (T)	7.90%	4.15
	meal	Guide to meal plans (I)		
	easy	Easy-to-use (U)		
	item	Customize meal plan (P)		
5	lose weight, help lose, pound, exercise diet, buddy activity, calorie deficit	Weight monitoring (T)	6.90%	4.73
	diet plan, healthy eat	Diet plans (I)		
	challenge	Encourage users on diets (E)		
6	track step, walk, mile, count, accurate, pedometer, count step, gps, distance	Walk tracker (T)	6.30%	4.03
7	drink water, water intake, hydrate, coffee, beverage, tea	Water drinking tracker (T)	5.60%	4.32
	notification, reminder, octopus	Encourage user drink water (E)		
8	track weight, bmi, weight loss, measurement, progress, goal, body fat, date, idea weight, moving everage	Weight monitoring (T)	5.00%	4.15
	chart, line graph, simple	Result visulization (U)		
	backup restore	Data backup and restore (S)		
9	keto diet, recipe, food, meal, low carb, eat, ingredient, macro, find, cook, option, list, tasty	Diet recipe (I)	4.90%	4.26
10	intermittent fast, timer, stage, fast plan, fast tracker, start/end	Fasting tracker (T)	4.20%	3.92
11	workout, exercise level, quick exercise, body, hard, video, variety, muscle, air squat, exercise equipment, feminine, great range, non-stop, workplace	Video guide to exercise (I)	3.80%	4.75
	beginner chanllenge, easy follow, goal, feel good	Encourage user by challenge (E)		
	routine	Exercise planning (P)		
12	barcode scanner, easy use, food item, interface, database, search product, barcode find, barcode feature, search brand, store brand, look food, helpful	Convenient food tracker using barcode scanner (U)	2.80%	4.31

Table 3. Results of Nutrition Apps

Results of Nutrition App Reviews

Table 3 summarized 12 emerged topics from the nutrition app reviews. The size of the topics varies from 2.80% to 24.40% percent. Topic 1 is about guide and planning of muscle build exercises, and sleep monitoring, which overlaps workout and meditation app subcategories. Topic 2 includes themes of food tracker issues, such as account login, disruptive advertisement, inconvenient editing the food items. Topic 3 focuses on the same issue as the Topic 6 in lifestyle app sub-category, which is the charge disputes. Topic 4 focuses predominantly on 4 different themes of calorie tracking apps. The first theme is the useful function of calorie tracking. The second theme is the useful guide to make health meal. The third theme is the easy-use experience. The last theme is the customizability of meal plan. Topic 5 focuses on exercise diet apps, including themes of weight loss monitoring function, diet planning, and encouraging users to complete the diet challenges. Topic 6 is about walk exercise tracker which is an overlapped topic. Topic 7 is about water drinking tracker, including themes of encouraging user to drink water more and track the total intake of water. Topic 8 focuses on visualization of weight loss results. Other themes in the topic are weight monitoring and backup data restore. Topic 9 is the rich information of diet recipe provided by apps. Topic

10 is the tracker of intermittent fasting. Topic 11 focuses mainly on the guide to exercise in the intuitive form of video. Exercise planning and using challenges to encourage users are other two themes in this topic. Topic 12 is the convenience of tacking food by barcode scanner.

The topic overlap of nutrition and workout sub-categories observed above implies that developers may design nutrition and workout features jointly when targeting clients to lose weight. We also noted that rich information of health food recipe, convenient barcode scanner for food, easy-to-use tracking, intuitive visualization and planning are expected by the user of nutrition app. Last, we found the pricing, charging and customer service related complains issues in this sub-category. In fact, this topic is a general negative factor appearing every sub-categories and worth attention

Topic No.	Key Terms	Theme Deduction	Percentage of Tokens	Average Ratings
1	guided meditation, mindfulness, practice, help, session, time, recommend, teacher, music, voice	Guided meditation (I)	19.70%	4.45
	free	Paywall Complaints (C)		
2	workout, exercise, yoga, breathe, stretch, beginner, weight, video, fitness, practice, trainer	Guide to exercise (I)	12.80%	4.52
	easy	Easy-to-use (U)		
3	fall asleep, help sleep, relax, drift away	Help sleep and relax (T)	12.00%	4.37
	sleep sound, story, wake, music, listen peaceful, night story, michelle sanctuary	Various peaceful sounds (I)		
	time, phone, wake alarm, night alarm	Sleep schedule (P)		
4	cancel subscription, payment, charge, free trial, refund, premium, money, card, service, even, try cancel, want, scam, version, paypal, billing, sign, customer service, credit card, bank	Price and Customer Service (C)	6.00%	2.29
5	white noise, sound, music, binaural beat, sleep, fan, listen, option	Various binaural beats (I)	8.40%	4.45
6	deep sleep, night, fall asleep, detect snore, time, sleep monitor, insomnia, track, rem, chart, bedtime routine, microphone, accuracy, rem sleep, sleep pattern, wake, use	Sleep monitoring (T)	8.00%	4.37
	free	Paywall Complaints (C)		
7	relax, fall sleep, music, sound, help, asleep, calm, meditation, listen, soothe	Calm music (A)	6.70%	4.74
8	reduce anxiety, help, panic attack, mental health, anxious stress, therapy, calm, depression, counsel	Reduce the anxiety (T)	5.10%	4.76
	offline mode	Paywall Complaints for offline mode (C)		
9	keep crash, open, update, load, try, time, reinstall, login, improve stability, account easy, able login, black screen, front screen, crash continuously, download install, wifi, issue open	Crash and login issues (S, O)	4.90%	3.87
	ability personalize, great customizable	customizable options (P)		
	easy use, simple, maneuver	easy-to-use (U)		
10	rain sound, thunderstorm, drink water, plant, sleep, cute, night, roof, rain wind, distant thunder	Rain sound (A)	4.00%	4.56
11	mix sound, volume control, different sound, save combination, option, custom, quality, timer, preset, arrangement, baby monitor, bitrate, edit sound, custom, play save, great UI, sound control, selection	Sound customizable (P)	3.70%	4.57
12	ad pop, sound, loud, intrusive, postcard, remove ad, screen ad, annoy, full screen, commercial, obnoxious, ad play, unskippable, interruption ad, ad ridiculous, hate ad, ad problem, ad begin, ad close, ad problem, ad uninstalled, pop everytime	Disruptive ads (U)	3.60%	4.05

Table 4. Results of Meditation Apps

Results of Meditation App Reviews

Table 4 summarized the 12 emerged topics from the meditation app reviews, whose size varies from 3.60% to 19.70% percent. Topic 1 consists of two themes focusing on guide to meditation through stillness or relax and complaints of the paywall. Topic 2 is about meditation through breathing or movement such as yoga and stretch exercises. The topic consists of two themes focusing on the guide to practice and the easy-to-use experience of the app. Topic 3 focuses on sleep improvement. Two themes of the topic are various relaxing sound and sleep cycle scheduling. Topic 4 focuses on charge dispute issues. Topic 5 focuses on various binaural beats. Topic 6 focuses mostly on sleep monitoring. Another theme of this topic is also complaints of the paywall. Topic 7 focuses on positive experience of calm music. Topic 8 is about relieving anxiety. Complaints of offline mode purchase are also one theme of this topic. Topic 9 consists of three themes about the functionality. The first theme are issues of account login and crash of the app. The second theme is great customizability. The last theme is maneuverability of the app. Topic 10 focuses on positive experience of rain sound. Topic 11 is about the popular function that allows sound customizable by users. Topic 12 focuses on disruptive ads.

We observed that the richness of sound or music is critical to the success meditation apps and mentioned most often by users. Developing meditation apps is also less challenging and may has higher success rate,

based on overall higher ratings of each topics comparing to other sub-categories. We also note that the app crash and account login are very common technique issues in meditation and other sub-categories.

Topic No.	Key Terms	Theme Deduction	Percentage of Tokens	Average Ratings
1	workout, exercise, body, day, beginner, kira, follow, fun, minute, need, move, level, variety, keep, start, help, jessica, body groove, melissa, vicky justiz, jiilian, dance, kira strokes, great range	Guide to dance workout (I)	17.00%	4.63
	chanllenge	Encourage user by challenge (E)		
	easy use	easy-to-use (U)		
2	cancel subscription, charge, free, pay, free trial, refund, money, email, month, customer service, sign, try cancel, payment, account, day, premium, membership, credit card, scam, bank, bill, debit, trial charge, renewal, impossible cancel	Price and Customer Service (C)	15.40%	1.95
3	different stretch, workout, exercise, day, body, start, gym, maternity, exercise library, tutorial, find pregnant, body, muscle, level, recommend, variety	Guide to gym exercises (I)	10.60%	4.64
	really easy, easy use	Easy-to-use (U)		
	track lift	Track lift exercises (T)		
	routine	Exercise planning (P)		
4	work, login screen, log, try open, reset password, crash, use, video, email click, ca, fix, update, network error, issue account, even, download, phone, load, try sign, reinstall, problem, credential, verify email, glitch, login issue, crash, update video, confirm email, data cache, try uninstalling, email address	Login issue and crash (O, S)	8.80%	2.20
5	food item, meal plan, calorie burn, eat, recipe, diet, scan, list, nutrition, intake, nutrient, carbs, ingredient, food diary, vitamin, macros, intake, database	Guide to meal plans (I)	8.70%	3.92
	track calorie, count, weight	Weight tracker (T)		
	barcode, scanner,	Convinent barcode scanner (U)		
6	yoga practice, pose, meditation, jessamyn, beginner, instruction, session, relax, teacher, video, asana, class	Guide to yoga (I)	8.10%	4.33
	voice, music	Motivated voice and music (E)		
7	plank workout, trainer, fitness, exercise, wods, equipment, program, amaze trainer, biceps back, alive inside, crossfit, elbow, best trainer, home	Guide to plank and crossfit (I)	8.00%	3.74
	feature parity, phone memory, workout crash, connect network, fix, update, load, version, bug annoy	Crash and network issues (S)		
8	bike, watch, connect, fitbit, device, pair, monitor, cadence sensor, sync, google, data, heart rate, garmin, strava, fit, resistance band	Connection issue with the sensors (M)	6.20%	3.34
9	tv, cast tv, chromecast, workout, cast option, would, make, screen, stream, connect tv, hearing, playlists, google tv, smart tv, roku	Connection issue with the TV cast (M)	5.10%	3.70
	music, voice, spotify music, audio, sound, phone volume	Pleasant music (A)		
10	track, gps, mile, distance, track progress, run, tracker, update, walk, plan, start , weight, gps, drain battery, run track, distance run, lane, last set, phone gps, , keep	Track walk exercise (T)	4.60%	4.09
	fasting, intermittent fast, stop fast, body fast	Track intermittent fast (T)		
11	lose weight, belly fat, pound, diet, help lose, exercise, reduce, recommend, reduce weight, extremely satisfied, follow diet, healthy diet	Exercise diet (I)	4.10%	4.72
12	amaze, easy use, simple, content user, amazing, excellent, beginner, awesome, need, english version, chinese version, miss package, hindi miss, 531 program, discipline great, thanks, respond question, helpful	Easy-to-use (U)	3.40%	4.72

Table 5. Results of Workout Apps

Results of Workout App Reviews

Table 5 summarized the 12 emerged topics from sub-category of workout app reviews. Topic 1 focuses on the dance fitness, including themes of guide to dance fitness, using challenges to encourage user and easy-to-use experience. Topic 2 is the charge disputes. Topic 3 focuses on the gym exercises, consisting of themes of guide to gym exercises, track lift exercises, planning and easy-to-use experience. Topic 4 in about the app crash and account login issues. Topic 5 overlaps the nutrition subcategories, consisting of themes of recipe, weight tracker and convenient barcode food scanner. Topic 6 focuses on yoga practice, including themes of guide to practice yoga and motivated voice and music. Topic 7 focuses on plank and cross-fitness, including themes of fitness guide and complaints of app crash and network issues. Topic 8 focuses on connection issue with the device sensors, such as smart watch, bike computer and band. Topic 9 is about the issue of TV cast. Plenty of great music is another theme in this topic. Topic 10 overlaps the lifestyle app sub-category about walk exercise and intermittent fasting tracker. Topic 11 overlaps the nutrition app sub-category about guide to exercise diet. Topic 12 is the general positive user experience.

From the above results, we note the fitness guide and encouragement provided by the workout apps is the core of successful design. Workout apps not only should provide extensive training information for certain fitness program, but also encourage users continue exercise by various persuasive techniques, including motivated music, virtual races, challenges, achievement system or even gamification. We also noted many

internet influencers and their channel, such as “Vicky Justiz”, “Kira Strokes” are frequently mentioned. Developers can introduce fitness influencers into app to enhance the coaching and encouragement features of the workout apps. Workout apps that require multi-device connection also face similar technique issue and complaints from users.

Topic No.	Key Terms	Theme Deduction	Percentage of Tokens	Average Ratings
1	doctor, health, appointment, insurance, need, care, information, medical, provider, test, card, schedule, appointment, workout, symptom, nurse, vaccination, hospital, qr, menopause, booster, caresource, gym, record, health care, record	Appointment scheduling issues (P)	40.90%	3.46
2	forgot password, log, fingerprint login, time, try log, sign, work, fingerprint, reset, change password, account, website login, reset username, code, wrong, every time, error, recognize device, account lock, get verification, reset username, sms, sso, try install, type password, authentication say, time log	Fingerprint login issues (O)	12.50%	1.45
3	refill prescription, pharmacy, pick order, medication, humana, rx, script, request refill, store, fill, prescription ready, call, reminder, delete, refill date, delivery, mile away, expire prescription, otc, cart, item	Pharmacy notification issues (P)	11.00%	2.67
4	open, upload, document, try, fix, crash, version, update available, need update, upload document, download, load, phone, screen, uninstalled, closing, prompt update, home address, update try, work fix, let update, throw, upload take, always shut, card information, cleaner, crash constantly, document like, download nothing, fail update, get open, hold hour, junk	Crash and document unavailable (S)	10.20%	1.77
5	claim, sync, samsung health, fitbit, health, information, fix, track, bcbs, connect, step count, step tracker, like error, pharmacy claim, sorry look, sync fitbit, access data, 3rd party, activity like, attempt say, benefit information, button, competent, detail enough, fsa claim, give permission, health connect, hire competent, count, issue, track, never connect	Claim information synchronization incorrect (M)	9.30%	2.04
	due inactivity,	Login issues (O)		
6	log, password, register, login, try, sign, account, cant, email, try reset, fix, enter, create, wrong, error, back signin, hard register, information correct, account try, able register, almost impossible, already register, create username, even create, almost impossible, extremely frustrating, login spin, message password	Login issues (O)	5.20%	1.47
7	meal, numi eat, calorie, weight loss, track, intake, log food, water	Food tracker (T)	4.78%	2.58
	nutrisystem food, journal, dietician, recipe, diet	Guide to food nutrition (I)		
	produce, snack, healthy	Price complaints of the produce sold (C)		
	scanner, scan	Convenient scanner (U)		
	item	Customize meal plan (P)		
8	email message, connect server, update, say, download, connect, login, new, server error, can not connect, new version, uninstalled, tell, message open, download enough, wait install, original version, server, can't connect	Connection issues with the server(S)	3.20%	1.41
9, 10	easy use, easy navigate, information easy, difficult accessibility, desktop site, navigate relevant, compass, convenient informative, abundance information, add echeck, amaze easy, fingertip great, everyting, user friendly, appreciate, convenient helpful, basic complaint, benefit record	Easy-to-use (U)	3.00%	4.62

Table 6. Results of Prescription Apps

Results of Prescription App Reviews

Table 5 Table 6 summarized the 9 emerged topics of sizes ranging from 3.00% to 40.90% percent from sub-category of prescription app reviews. With the predominately largest size, topic 1 focuses on the appointment scheduling issues. Topic 2 focuses on complaints of login issues. Topic 3 focuses on the missing notification of notification from pharmacy. Topic 4 is about complaints of app crash and unavailable health document. Topic 5 is about complaints of incorrect claim information and data synchronization, login issues due to inactivity. Topic 6 focuses on account login issues due to various reasons. Topic 7 overlaps the nutrition app sub-category, consisting of similar themes of food tracker, guide to nutrition, complaints of the price of the produce sold, convenient barcode scanner and meal customization. Some prescription apps advertise other nutrition apps inside their apps. Topic 8 is about the complaints of issues of connecting server. Topic 9 focuses on easy navigating of the apps.

We noted the prescription apps received overwhelmed negative feedback and significant lower ratings compared to other sub-categories. The complaints cover various techniques issues including account login, connecting to the server, data synchronization, crash, notification lost. The low quality may be caused by the fragmentation of developing, that is, each institution developed their own apps with limited budget and cause the low quality of those apps. A unified app severed as a third-party platform for all institutions may solve the problem. The easy navigation is the most expected from users.

Discussion

We plotted the tripartite graph (Figure 5, 7) to visualize the relationships between theme groups and emerged topics from the 5 sub-categories. Five dots in the left column represent the 5 sub-categories. The dots in the middle column represent the emerged topics from each sub-category, with the line of the same colour of sub-category dot connected. The y-coordinate values of topic dots in the middle are the average ratings of each topic. The dots on the right column are the theme groups. We introduced the software quality metrics and their definitions proposed by (Gaffney Jr 1981) and aligned some of them (Table 7 Column 1) with the most related theme groups (Table 7 Column 3) from our results. We ignore the metrics such as portability or maintainability, since the users have no access to the source code. We also introduced the WisCom topic modelling of mobile app reviews based on LDA proposed by (Fu et al. 2013) (Table 7 Column 2, Table 8 Column 1). They are aligned to theme groups by comparing keywords.

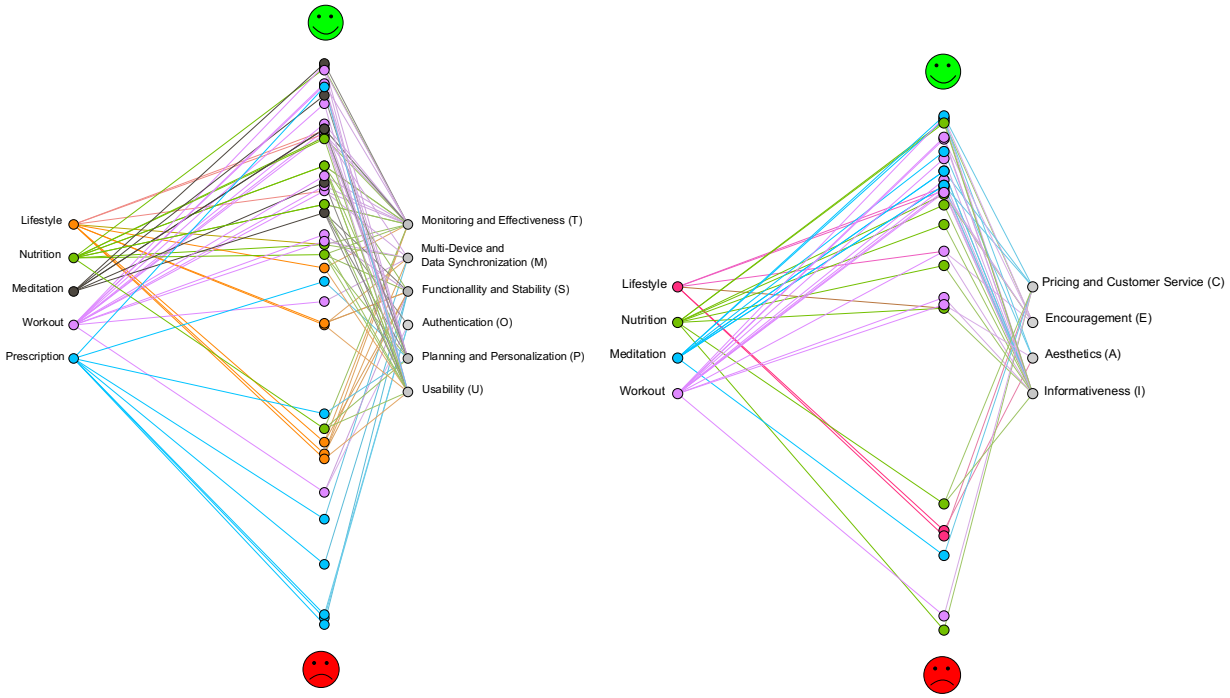

Figure 5. Relations between Issues and Subcategories

After reviewing the result, we suggest that developers should thoroughly exam the designs under metrics mentioned in the following. The correctness metric means the extent to which a program fulfils the user, which means effectiveness of monitoring and tracking (T) in app design. The tracking accuracy of sleep, food and weight received relatively positive feedback, however, the accuracy of health signals, location, walking steps is often complained by the users. The interoperability metric means the effort required to couple one system with another. Tracking apps work under multiple device and sensor environment (M) face the challenge of connection and synchronization issues, such as connected to the smart watch, cadence sensor and sync up with google fit account. The reliability metrics means the extent to which a program satisfies its specifications. This often means functionality and stability (S) issues of mobile apps, especially critical errors that cause app crash. It frequently happens when users log in account, connect to the network, restore the backup data, sync with the server and update it to new version. The integrity metrics means the extent to which access to software or data by unauthorized persons can be controlled. Account login is one of the most complained issues already discussed above. The flexibility metrics means the effort required to modify an operational program, which means the planning and personalization (P) for app users, such as exercise planning. Except text editing (Topic 2 in Nutrition apps) and appointment scheduling issues (Prescription apps), mobile apps perform very well on flexibility in general, which may be contributed by the mature UI testing technology today. Finally, the usability metric means the effort required to learn, operate, prepare input and interpret output of programs, which means usability of apps. Most positive feedback are easy-to-use. Weight tracker result visualization and food barcode scanner are convenient

features mentioned. Although disruptive advertisements affect useability negatively, users seem more tolerate to them compared to charge disputes.

Software Quality Metrics		WisCom-topic Model of Mobile Apps		Health and Fitness App Quality Metrics	
Metric	Definition	Topic	Key words	Metric	Examples
Correctness	Extent to which a program fullfills the user's mission objectives	Accuracy	find, location, search, info, useless, data, way, list, sync, wrong	Monitoring and Effectiveness (T)	fitness tracker, track or record activities, monitor health signal, help lose weight or fall asleep
Interoperability	Effort required to couple one system with another	Compatibility	galaxy, battery, support, off, droid, nexus, compatible, install, samsung, worked	Multi-Device and Data Synchronization (M)	Issues under multi-device enviroment, pair and data synchronization
Reliability	Extend to which a program satisfies its specifications	Stability	closes, close, load, every, crashes, keeps, won, start, please, closing	Functionallity and Stability (S)	no crash or other severe errors, function properly after upgrade, version
Integrity	Extent to which access to software or data by unauthorized persons can be controlled	Connectivity	log, error, account, connect, login, connection, sign, let, slow, website	Authentication (O)	user account management, access user's account and profile data successfully
Flexibility	Effort required to modify an operational program	Picture	pictures, picture, pics, camera, save, wallpaper, see, photes, upload, pic	Planning and Personalization (P)	build exersise plan, schedule appointment with health care provider, customize meal plan
Usability	Effort required to learn, operate, prepare input, and interpret output of program	Spam	ads, notification, spam, bar, notifications, adds, annoying, many, pop, push	Usability (U)	simple and intuitive interface, frictionless user experience, easy to navigate and use, minimal interuption, no spam

Table 7.

WisCom-topic Model of Mobile Apps		Health and Fitness App Quality Metrics	
Topic	Key words	Metric	Examples
Cost	free, money, buy, pay, paid, refund, want, back, bought, waste	Pricing and Customer Service (C)	charging dispute, payment experience, price, paywall, membership, premium, in-app purchasing
Telephony	uninstall, want, need, send, message, delete, let, contacts, calls, off	Encouragement (E)	persuasive features including gamification and social encouragement to increase physical exercise.
Attractiveness	boring, bad, stupid, waste, dont, hard, make, way, graphics, controls	Aesthetics (A)	beautiful design of user interface, nice bgm
Media	video, sound, watch, videos, songs, audio, sounds, hear, record, anything	Informativeness (I)	informativeness and extensiveness of the content, variety of tutorial and guidance, influencer-drivern

Table 8.

The WisCom-topic study argues that the ten most common topics are emerged from apps of all categories. Four theme groups of our results are related to WisCom-topics (Table 8). The first theme group is the pricing and customer service (C) issue, which is a common complaint in every sub-categorizes. To improve the payment processes, developers may consider use API provided by third-party fintech company such as Swipe. The developer may also consider adjust the price or even switch monetization from purchase to advertisement, since users relatively tolerate it. The theme group encouragement (E) is about persuasive techniques which are discussed in Section 4.4. Those techniques have significant positive effects on nutrition and workout apps. The aesthetics (A) plays particularly important role in meditation apps. Beautiful smart watch face is very popular, but it requires efforts to allow users customize themes smoothly. The fourth theme group is informativeness (I). Today, mobile apps often server as media platforms, therefore, the rich and informative content is important as the examples of meditation, workout, nutrition apps show. The four new metrics we discussed above extended the design of focus from software quality to overall values brought to the customers.

Sub-category	Apps as search seeds	Number of apps	Number of reviews	Minimal size of cluster
Lifestyle	Fitbit, VeryFitPro, Strava, Runkeeper	410	38592	100
Nutrition	Lifesum, YAZIO	288	35035	70
Meditation	Calm, BetterSleep	225	21161	50
Workout	Smartabase Athelete, Fiton, Peloton	334	28130	70
Prescription	Humana, Go365 for Humana,	95	8135	30

Table 9. Statistics

Conclusion

We are now face a highly competitive and lucrative mobile app market ever growing. This paper examined the emerged topics from health and fitness app reviews using AI-based approach. We deducted the themes of sub-category for specific app design guidelines. We also discussed quality metrics related themes and provided all error-prone scenarios which should be thoroughly tested to assure the quality of apps. The discussion of new theme groups extends the quality metrics with new customer values. All above guidelines from deep understanding of the nature and characteristics of customer feedbacks lay the foundations for the design of next popular app. The whole framework can quickly be applied to any new category of apps and improves the chance of successful design.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. 2001. "On the Surprising Behavior of Distance Metrics in High Dimensional Space," *International conference on database theory*: Springer, pp. 420-434.

Allaoui, M., Kherfi, M. L., and Cheriet, A. 2020. "Considerably Improving Clustering Algorithms Using Umap Dimensionality Reduction Technique: A Comparative Study," *International Conference on Image and Signal Processing*: Springer, pp. 317-325.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. 1999. "When Is "Nearest Neighbor" Meaningful?," *International conference on database theory*: Springer, pp. 217-235.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of machine Learning research* (3:Jan), pp. 993-1022.

Chen, N., Lin, J., Hoi, S. C. H., Xiao, X., and Zhang, B. 2014. "Ar-Miner: Mining Informative Reviews for Developers from Mobile App Marketplace," *Proceedings of the 36th international conference on software engineering*, pp. 767-778.

Chen, R., Wang, Q., and Xu, W. 2019. "Mining User Requirements to Facilitate Mobile App Quality Upgrades with Big Data," *Electronic Commerce Research and Applications* (38), p. 100889.

Dalpiaz, F., and Parente, M. 2019. "Re-Swot: From User Feedback to Requirements Via Competitor Analysis," *International working conference on requirements engineering: foundation for software quality*, pp. 55-70.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*).

Fang, X., and Zhan, J. 2015. "Sentiment Analysis Using Product Review Data," *Journal of Big Data* (2:1), pp. 1-14.

Flory, L., Osei-Bryson, K.-M., and Thomas, M. 2017. "A New Web Personalization Decision-Support Artifact for Utility-Sensitive Customer Review Analysis," *Decision Support Systems* (94), pp. 85-96.

Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., and Sadeh, N. 2013. "Why People Hate Your App: Making Sense of User Feedback in a Mobile App Store," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1276-1284.

Gaffney Jr, J. E. 1981. "Metrics in Software Quality Assurance," *Proceedings of the ACM'81 conference*, pp. 126-130.

Goldberg, D. M., and Abrahams, A. S. 2022. "Sourcing Product Innovation Intelligence from Online Reviews," *Decision Support Systems*), p. 113751.

Grootendorst, M. 2022. "Bertopic: Neural Topic Modeling with a Class-Based Tf-Idf Procedure," *arXiv preprint arXiv:2203.05794*).

Gu, X., and Kim, S. 2015. "" What Parts of Your Apps Are Loved by Users?"(T)," *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 760-770.

Guzman, E., and Maalej, W. 2014. "How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews," *2014 IEEE 22nd international requirements engineering conference (RE)*, pp. 153-162.

Ha, S., and Geum, Y. 2022. "Identifying New Innovative Services Using M&a Data: An Integrated Approach of Data-Driven Morphological Analysis," *Technological Forecasting and Social Change* (174), p. 121197.

Johann, T., Stanik, C., Maalej, W., and others. 2017. "Safe: A Simple Approach for Feature Extraction from App Descriptions and App Reviews," *2017 IEEE 25th international requirements engineering conference (RE)*, pp. 21-30.

Lee, G., and Raghu, T. S. 2014. "Determinants of Mobile Apps' Success: Evidence from the App Store Market," *Journal of Management Information Systems* (31:2), pp. 133-170.

Li, Q., Zeng, D. D., Xu, D. J., Liu, R., and Yao, R. 2020. "Understanding and Predicting Users' Rating Behavior: A Cognitive Perspective," *INFORMS Journal on Computing* (32:4), pp. 996-1011.

Li, Y.-M., Chen, H.-M., Liou, J.-H., and Lin, L.-F. 2014. "Creating Social Intelligence for Product Portfolio Design," *Decision Support Systems* (66), pp. 123-134.

Liu, X., Ai, W., Li, H., Tang, J., Huang, G., Feng, F., and Mei, Q. 2017. "Deriving User Preferences of Mobile Apps from Their Management Activities," *ACM Transactions on Information Systems (TOIS)* (35:4), pp. 1-32.

Liu, Y., Jiang, C., and Zhao, H. 2018. "Using Contextual Features and Multi-View Ensemble Learning in Product Defect Identification from Online Discussion Forums," *Decision Support Systems* (105), pp. 1-12.

Liu, Y., Jiang, C., and Zhao, H. 2019. "Assessing Product Competitive Advantages from the Perspective of Customers by Mining User-Generated Content on Social Media," *Decision Support Systems* (123), p. 113079.

Liu, Z., Qin, C.-X., and Zhang, Y.-J. 2021. "Mining Product Competitiveness by Fusing Multisource Online Information," *Decision Support Systems* (143), p. 113477.

Maalej, W., and Nabil, H. 2015. "Bug Report, Feature Request, or Simply Praise? On Automatically Classifying App Reviews," *2015 IEEE 23rd international requirements engineering conference (RE)*, pp. 116-125.

Mabey, B. 2018. "Pyldavis: Python Library for Interactive Topic Model Visualization," *Port of the R LDAvis package*).

Malik, H., Shakshuki, E. M., and Yoo, W.-S. 2020. "Comparing Mobile Apps by Identifying 'Hot'features," *Future Generation Computer Systems* (107), pp. 659-669.

McInnes, L., Healy, J., and Astels, S. 2017. "Hdbscan: Hierarchical Density Based Clustering," *J. Open Source Softw.* (2:11), p. 205.

McInnes, L., Healy, J., Saul, N., and Großberger, L. 2018. "Umap: Uniform Manifold Approximation and Projection," *Journal of Open Source Software* (3:29), p. 861.

Mortenson, M. J., and Vidgen, R. 2016. "A Computational Literature Review of the Technology Acceptance Model," *International Journal of Information Management* (36:6, Part B), pp. 1248-1259.

Reimers, N., and Gurevych, I. 2019. "Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks," *arXiv preprint arXiv:1908.10084*).

Rob van der Meulen, J. R. 2014. "Gartner Says Less Than 0.01 Percent of Consumer Mobile Apps Will Be Considered a Financial Success by Their Developers through 2018." from https://www.gartner.com/en/newsroom/press-releases/2014-01-13-gartner-says-less-than-one-tenth-percent-of-consumer-mobile-apps-will-be-considered-a-financial-success-by-their-developers-through-2018

Shah, F. A., Sabanin, Y., and Pfahl, D. 2016. "Feature-Based Evaluation of Competing Apps," *Proceedings of the International Workshop on App Market Analytics*, pp. 15-21.

Shah, F. A., Sirts, K., and Pfahl, D. 2019. "Using App Reviews for Competitive Analysis: Tool Support," *Proceedings of the 3rd ACM SIGSOFT International Workshop on App Market Analytics*, pp. 40-46.

Sievert, C., and Shirley, K. 2014. "Ldavis: A Method for Visualizing and Interpreting Topics," Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 63-70.

Singh, A., and Tucker, C. S. 2017. "A Machine Learning Approach to Product Review Disambiguation Based on Function, Form and Behavior Classification," *Decision Support Systems* (97), pp. 81-91.

Thakur, N., Reimers, N., Daxenberger, J., and Gurevych, I. 2021. "Augmented Sbert: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 296-310.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention Is All You Need," *Advances in neural information processing systems* (30).

Verkijika, S. F., and Neneh, B. N. 2021. "Standing up for or Against: A Text-Mining Study on the Recommendation of Mobile Payment Apps," *Journal of Retailing and Consumer Services* (63), p. 102743.

Wallach, H. M. 2006. "Topic Modeling: Beyond Bag-of-Words," *Proceedings of the 23rd international conference on Machine learning*, pp. 977-984.

Zheng, X., Zhu, S., and Lin, Z. 2013. "Capturing the Essence of Word-of-Mouth for Social Commerce: Assessing the Quality of Online E-Commerce Reviews by a Semi-Supervised Approach," *Decision Support Systems* (56), pp. 211-222.

Zhu, D., Lappas, T., and Zhang, J. 2018. "Unsupervised Tip-Mining from Customer Reviews," *Decision Support Systems* (107), pp. 116-124.