Original Research Article

Validation of Some Health Fitness Apps Using Users' Reviews

ABSTRACT

With the increase in the number of Health Fitness Applications (Apps) available for free, there is a growing concern as to whether these apps actually help individuals achieve personal fitness. This research developed a system to validate three Health Fitness Apps before user download using user reviews.

Sentiment Analysis as the application of natural language processing, computational linguistics, and text analytics was used to identify and classify subjective opinions in the reviews of three most commonly used Health Fitness Applications; Samsung Health, Google Fit and Home Workout. Analysis showed that the Home Workout Fitness Application garnered a total of 99.9% Positive Reviews and can therefore be said to be the most effective of the three Apps considered, followed by Google Fit Fitness Application with a total of 37.4% Positive Reviews and Samsung Health Fitness Application recorded the most Negative Reviews of 96.6%.

Keywords: Health Apps, Fitness Apps, Sentiment Analysis, Opinion mining, users' reviews, App validation.

1. INTRODUCTION

There is no doubt that Technology and the Internet have redefined the human social norms. Mobile Apps have invariably become the order of the day to make life easier for people, most especially because of the ubiquitous nature of mobile devices. Mobile Apps are available in all the industries, most especially the banking, health care and e-commerce sectors. A major feature of most Apps is the feedback system where users give comments on the Apps majorly based on their use and observations of the Apps. These reviews have helped in generation of profit for the developers, giving bug reports, request for new features, documentation of experience to analysts [1] and designers [12]. They give information related to products, services, organizations, individual's issues, events, satisfaction or dissatisfaction with the features or business relevant information. New apps are rolling out every day with technical information available in the description; and ordered in terms of latest reviews, ratings and download strategy [16]. According to a report from Forbes Unlimited articles in 2020, there are about 8.9 million mobile Apps. Reviews and observations from the entire App market may prove to be robust but manual analysis has become almost impossible since the app market stores a large number of reviews that takes longer computations and efforts, the quality of reviews vary tremendously from helpful advice to a bad advice and filtering the negative and positive comments in the reviews are sometimes tricky.

Sentiment Analysis helps to mine the people's opinions, sentiments, behaviors, emotions, appraisals and attitudes towards products or services, issues or events and topics [21]. There are three types of people's opinions namely; positive, negative and neutral, which identify the entire knowledge of the domain. It is an integral part of the Natural Language Processing (NLP) and helps in text mining and information retrieval. In recent years, it has extended to fields like marketing, finance, political science, communications, health science etc.

Machine Learning (ML) based techniques as well as lexicon-based methods are used in Sentiment Analysis [9]. Lexicon based approach is an approach that considers the semantic order of the words and doesn't include labelled data, dictionary is created manually and includes words and phrases in a document. Sentiment Analysis through Machine

Learning approach deals with labelled data and helps to create models using supervised learning algorithms such as, Na¨ıve Bayes (NB), Support Vector Machine (SVM) and K-nearest neighbor (KNN) [10].
However, the influx of fake and sub-standard Health Fitness Applications has been on the rise, hence the need to develop methods of validating which of these Health Fitness Apps perform according to specifications using the reviews of previous users. This will help potential users of these Apps to make informed choices about them.

2. LITERATURE REVIEW

Fitness is the state of being physically fit and healthy through proper exercise, diet, and sleep habits [11]. Physical inactivity is a major cause of lack of fitness; according to the World Health Organization, approximately 31% of adults are insufficiently active, and approximately 3.2 million deaths each year are due to insufficient physical activity (WHO, 2014). Given that healthcare providers see their patients for only a brief moment of time, it is a challenge to motivate them to pursue health and fitness in between visits. Internet-based systems have been used to motivate patients to pursue fitness and seem to improve compliance [18]. A review of internet based physical activity interventions noted that 61% of studies reported significant increases in physical activity [8]. With the rapid increase of smartphones used in recent years, a new way of monitoring and motivating patients to engage in health and fitness is taking shape.
Health Assistance Applications can be viewed as mobile application program that offer health-Fitness services on smart phones, tablets, PCs and other communications devices. These Apps are linked to the fitness sector, revolutionizing the ways of doing physical activity and the relationships between fitness providers and consumers [14].
App Annie's State of Mobile 2021 Report estimated that more than 71,000 Health and Fitness Apps were launched in 2020 and this trend continued upward in 2021. The popularity of Fitness Apps came with its advancing feature of integrating wearable technology which helps users to mange their fitness programs [13].
Three widely used Fitness applications are used for the purpose of this project, namely: SAMSUNG HEALTH, GOOGLE FITE and HOMEWORKOUT.

2.1 Samsung Health

Samsung Health (originally S Health) is a free application developed by Samsung that serves to track various aspects of daily life contributing to wellbeing such as physical activity, diet, and sleep. Launched on 2 July 2012, [15] with the new Samsung smartphone, the Galaxy S3, the application was installed by default only on some smartphones of the brand. Presently, the application is installed by default on some Samsung smartphone models and it could also be downloaded from the Samsung Galaxy Store. The app changed its name from S Health to Samsung Health on 4 April 2017, when it released version 5.7.1. Some of its features include; Setting goals or using the goals suggested by the app to improve its results, Pedometer, Weekly summaries of the main features, Activity tracking, Dietary monitoring, Weight tracking, Sleep monitoring, monitoring of water consumption and blood sugar.

2.2 Google Fit

"Google Fit is a health-tracking platform developed by Google for the Android operating system, Wear OS and Apple Inc.'s iOS. It is a single set of APIs that blends data from multiple apps and devices" [2]. Google Fit uses sensors in a user's activity tracker or mobile device to record physical fitness activities (such as walking, cycling, etc.), which are measured against the user's fitness goals to provide a comprehensive view of their fitness.
"In August 2018, Google announced a revamp to its Android Fit platform which adds activity goals based on activity recommendations from the American Heart Association and the World Health Organization". [5]. "The updates are meant to help Fit better provide metrics for activities other than walking, and encourage users to engage in activities that will raise the heart rate without necessarily requiring a trip to the gym"[19].
"In April 2019, Google announced Google Fit for iOS offering similar experience to its Android counterpart. Google Fit for iOS used Apple Health, Nike Run Club, Headspace or connected device such as Apple Watch or Wear OS smartwatch connected to user device" [3,4].

2.3 Home Workout

Home Workout provides daily workout routines for all main muscle groups. In just a few minutes a day, one can build muscles and keep fitness at home without having to go to the gym with no equipment or coach needed, all exercises can be performed with just the body weight. The App has workouts for abs, chest, legs, arms and butt as well as full body workouts. All the workouts are designed by experts. None of them needs any equipment, so there's no need to go to the gym. Even though it just takes a few minutes a day, it can effectively tone muscles and help get six pack abs at home. The warm-up and stretching routines are designed to make sure exercise is done in a scientific way. With animations and video guidance for each exercise, one can make sure one uses the right form during each exercise. Some of its features include; Warm-up and stretching routines, Records training progress automatically, tracking weight trends, customizing

workout reminders, Detailed video and animation guides, Lose weight with a personal trainer and Share with friends on social media.

2.4 Related Works

[17] carried out a research to discover users' perception of health and fitness apps with the Extended Unified Theory of Acceptance and Use of Technology (UTAUT2) Model. The research was conducted with college-aged smartphone users at a Midwestern university in the United States and it was discovered that performance expectancy, hedonic motivations, price value, and habit were significant predictors of users' intention of continued usage of health and fitness apps.
[7] conducted a study on the use of Smartphone Apps for patient's health and fitness and introduced evidence that apps can better help patients reach their health and fitness goals. He also discussed what features to look for in an health and fitness App with an overview of popular health and fitness apps.
[20] developed a method to identify comparative reviews for mobile apps from an app market. 5 million reviews from Google Play were used and manual assessments on 900 reviews were done. The method was able to identify opinions accurately and provide meaningful comparisons between apps.
[6] proposed a new approach for Software Requirements Evolution using user review mining. The user review data was collected by crawling on the app review page and analyzed to check user satisfaction. It analyzed user with a machine learning method. A case study was conducted with a non-face-to-face video conferencing app and was concluded that Software improvement through user review mining contributes to the user lock-in effect and extends the life cycle of the software.

3. METHODOLOGY

Validation of Health Fitness Applications is aimed at determining which of the Health Fitness Apps on Google Playstore is more effective and performs its developer's specifications by making use of reviews of previous users. Three (3) of the most popular Health Fitness Apps were used as a case study (Google Fit, Samsung Health and Home Workout). The research was carried out using Sentiment Analysis.

3.1 Data Collection

The dataset contains the reviews scraped from Google playstore for the three Health Fitness Applications that are being considered, namely; Google Fit, Samsung Health and Home Workout (Health Assistance Apps). This dataset was scraped using a python library called google-play-scraper. A total of 50,000 Reviews were considered with 25,000 used for Training the system and the remaining 25,000 were used for Testing (Numbers of Reviews per each of the Health Fitness Applications are diffferent but the same number of Testing results were considered i.e 5000 each).

3.2 Data Preprocessing

To preprocess the data for further use, the following operations were carried out: Removal of Unlabeled Reviews, Train and Test dataset Split, Transforming datasets to pandas dataframe objects, Creating Input Sequences and Tokenization. However, the accuracy of the model also depends on the cleanliness of the data.

3.3 Manual Labelling

The reviews were manually labelled and grouped into Positive and Negative categories. Example as shown in the table below:

Table 1: Reviews were manually labelled and grouped into Positive and Negative categories

Reviews	Category
Really like this app (Google Fit)	Positive
Boring (Google Fit)	Negative
I have been using this app for several years now. Had to get a new phone. Now I can not track my stress level anymore. And I went from being on level 27 for my steps and now I'm back to level 2. All my numbers loaded but the level never carried over. (Samsung Health)	Negative
Excellent app for sports and fitness where we can have access to a lot of things related to our physical activities. Its nice to see an app like this existing. Its really great. (Samsung Health)	Positive
I'm old , fat and flabby and need something to get me started. This	Positive

seems to be it. It makes me workout with planned exercise. (Home Workout)	

3.4 Data Modelling

A function was created called “create_model”. The BERT tf module was loaded (to extract the computation graph). A single new layer was then created (this was trained to adapt BERT to the sentiment task (classifying reviews as positive or negative)). Within the function creation, the label of the sentiments in the training dataset was one hot encoded (i.e conversion of categorical values to numerical values) to prepare it for the model building. Thereafter, dynamism was applied in such a way that if the model is working on prediction, the predicted labels and probabilities will be returned but if it is working on training and evaluation, the loss will be calculated between the actual labels and predicted labels, then the loss, predicted labels and probabilities will be returned.
To build the function, a general "model_builder_fn()" function was defined, with a sub-function called "model_fn()". The sub-function "model_fn()" was built in such a way that the function sets the conditions for when prediction is to be done and when training and evaluation should be done. A second sub-function called "metric_fn()" was created to calculate the evaluation metrics for when the condition for evaluation is met. These conditions were set in such a way that when the mode is not set to "PREDICT", either the training (the mode has to be set to "TRAIN", the model will be trained and then the loss and training steps are returned) or the evaluation (the mode has to be set to " EstimatorSpec", the evaluation is done using evaluation accuracy, f1_score, AUC score, precision, recall, true_positives, true_negatives, false_positives, false_negatives as evaluation metrics) depending on the modes set.
An input builder function was created after which the model was trained using the training features dataset. The training took about 14 minutes as Google Colab's GPU was used. To check the performance of the model built, the model was applied to the test dataset and it produced an evaluation accuracy of 89% and loss of 0.45.

3.5 Sentiment Prediction

From the built model, a function “getprediction” was created. This function gets sentences as input/argument and then returns the sentences, the prediction probabilities and then the labels (0 for negative, 1 for positive) and from the “getprediction()” function, another function, “getsentiment()”, was created. A variable “review” was used to store the data read into the workspace using “pd.readcsv()”, the data was cleaned and passed into the function “getprediction()”, the result of the prediction was saved in a CSV as the output.

This function was then applied to every one of the review datasets of each Health Assistance App (Google Fit, Samsung Health and Home Workout). The model built has an average accuracy of 89%. Zoho Analytics was used for the visualization of results.

3.6 Evaluation Metrics

The evaluation metrics used are:

Accuracy: It's the magnitude relation of the variety of correct predictions to the whole variety of input samples. The closer it is to 1, the better. The formula is shown in eqtn 4.1:

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made} \qquad \text{eqtn 4.1}$$

F1_score: F1-Score is the harmonic mean of precision and recall values for a classification problem. The closer it is to 1, the better. The formula for F1-Score is shown in eqtn 4.2 below:

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \qquad \text{eqtn 4.2}$$

AUC: AUC (Area Under Curve)-ROC (Receiver Operating Characteristic) is a performance metric, based on varying threshold values, for classification problems. As the name suggests, ROC is a probability curve and AUC measures the separability. In simple words, AUC-ROC metric will reveal the capability of the model in distinguishing the classes. The higher the AUC, the better the model.

True Positives: This shows the number of reviews that the model predicted belong to a specific class and they actually belong to the class.The higher it is, the better.

True Negatives: This shows the number of reviews that the model predicted do not belong to a specific class and they actually do not belong to the class. The higher it is, the better.

False Positives: This shows the number of reviews the model classified under a specific category whereas those reviews do not belong in that category. The lower it is, the better.

False Negatives: This shows the number of reviews the model did not classify under a specific category whereas those reviews actually belong to that category. The lower it is, the better.

Precision: Precision is the total number of correct classifications returned by the model. The closer it is to 1, the better. It is the ratio of true positives to the sum of the true positives and false positives. It is shown in eqtn 4.3 below:

$$\text{Precision} = \frac{TP}{(TP+FP)} \qquad \text{eqtn 4.3}$$

Recall: Recall tells us the rate of the predicted positive reviews out of all actual positive reviews. The closer it is to 1, the better. It is the ratio of the true positives to the sum of the true positives and false negatives, as shown in eqtn 4.4:

$$\text{Recall} = \frac{TP}{(TP+FN)} \qquad \text{eqtn 4.4}$$

4. RESULTS AND DISCUSSION

The evaluation of the system gave the metrics scores below and shows that the system is valid and trust worthy:
AUC: 0.898

Accuracy: 0.898

F1_Score: 0.896

False Negatives: 282.0

False Positives: 229.0

Precision: 0.906

Recall: 0.887

True Negatives: 2281

True Positives: 2208

The scores from the metric are gotten from the BERT model.

BERT (Bidirectional Encoder Representations from Transformers) model is an open source machine learning framework for natural language processing (NLP). It is designed to help computer understand the meaning of ambiguous language in text using surrounding text to establish context. BERT is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. (In NLP, this process is called attention).

How BERT Works

The goal of any given NLP technique is to understand human language as it is spoken naturally. In BERT's case, this typically means predicting a word in a blank. To do this, models typically need to train using a large repository of specialized, labeled training data. This necessitates laborious manual data labeling by teams of linguists.
BERT, however, was pre-trained using only an unlabeled, plain text corpus (namely the entirety of the English Wikipedia, and the Brown Corpus). It continues to learn unsupervised from the unlabeled text and improve even as its being used in practical applications (i.e Google search). Its pre-training serves as a base layer of "knowledge" to build from. From there, BERT can adapt to the ever-growing body of searchable content and queries and be fine-tuned to a user's specifications. This process is known as transfer learning.

4.1 Samsung Health App

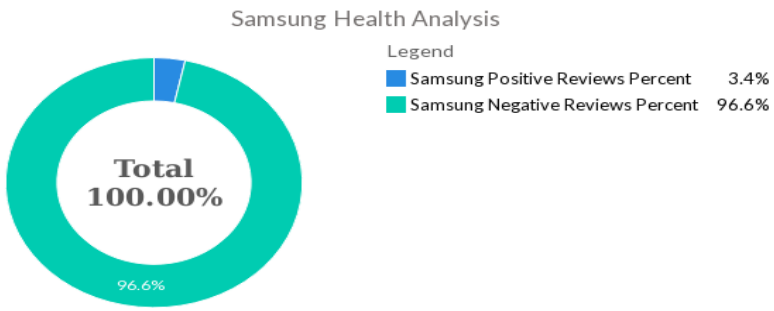


Fig 1: Sentiment Percentage Analysis of Samsung Health App

The chart ring in Fig 1 above shows the percentage of positive reviews to negative reviews in for the Samsung Health App. According to the chart, there is a high percentage of negative reviews for the App while the positive reviews are barely noticeable. Fig 2 below shows the sentiment count for the Samsung Health App

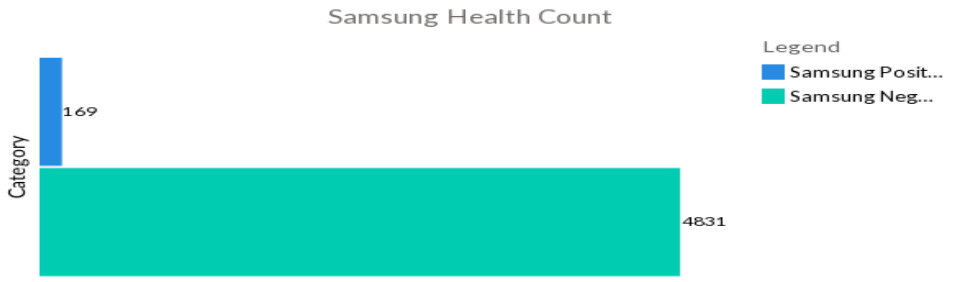


Fig 2: Sentiment Count of Samsung Health App

The green bar represents the negative reviews count and the blue bar represents the positive reviews. It is obvious that there are more negative reviews than positive reviews. According to the analyzed user reviews, this health app is not reliable.

4.2 Google Fit App

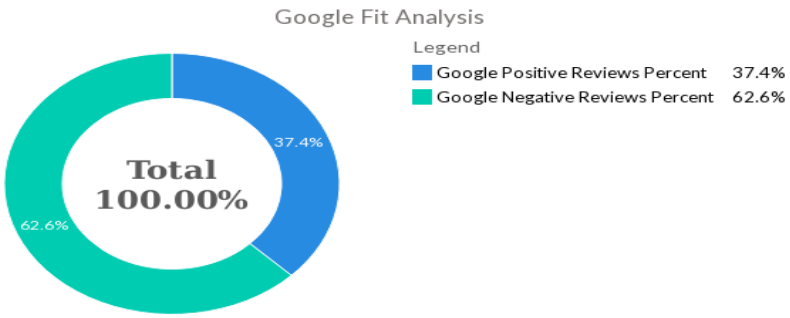


Fig 3: Sentiment Percentage Analysis of Google Fit App

The chart ring in Fig 3 above shows the percentage of positive reviews to negative reviews for the Google Fit App. According to the chart, there is a higher percentage of negative reviews for the App than the positive reviews. Fig. 4 shows the sentiment count for the App. The green bar represents the negative reviews count and the blue bar represents the positive reviews. It is also obvious that there are more negative reviews than positive reviews.

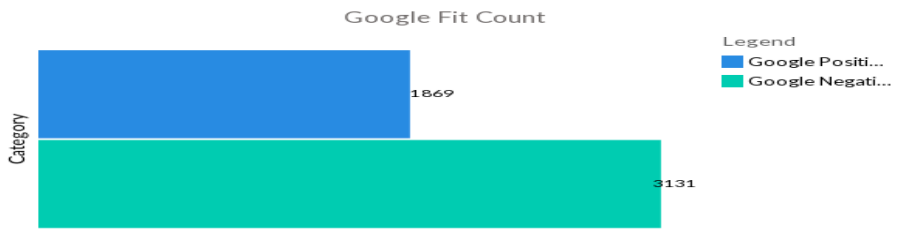


Fig 4: Sentiment Count of Google Fit App

4.3 Home Workout App

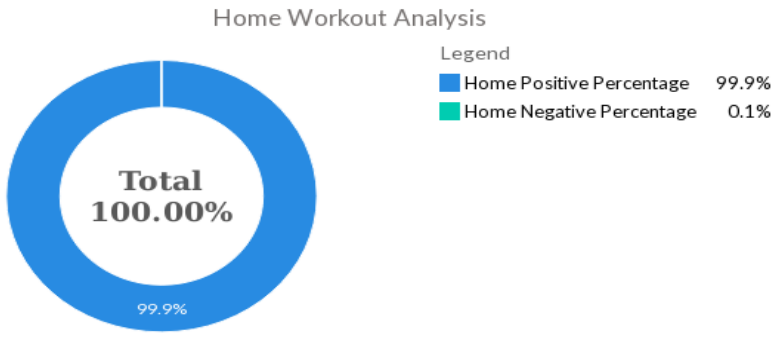


Fig 5: Sentiment Percentage Analysis of Home Workout App

The chart ring in Fig 5 above shows the percentage of positive reviews to negative reviews for the Home Workout App. According to the chart, there is an almost 100 percentage of positive reviews with the percentage negative reviews being almost non existent. Fig. 6 below shows the sentiment count for the App. The green bar represents the negative reviews count and the blue bar represents the positive reviews.

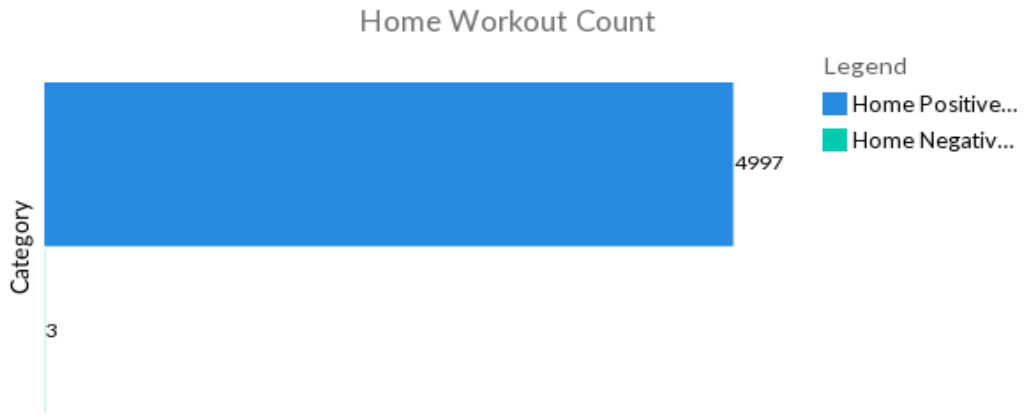
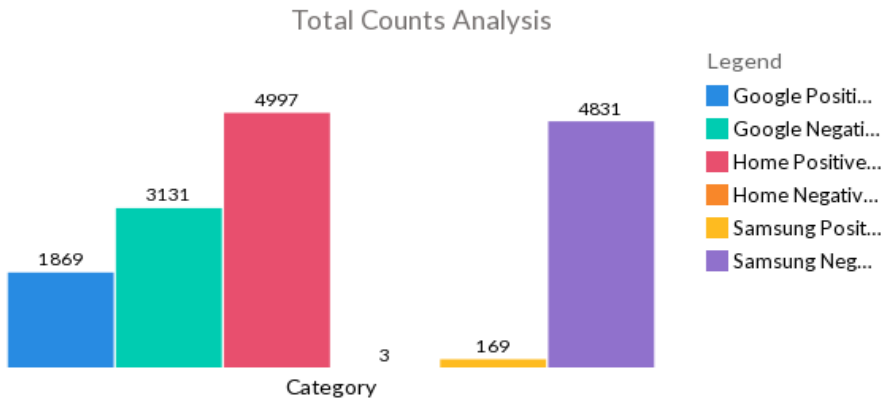


Fig 6: Sentiment Count of Home Workout App

The Home Workout app can therefore be said to be reliable and fit for its purpose.

4.4 Total Sentiment Counts

The total sentiment counts for the three (3) Health fitness Apps considered is presented in the Fig. 7 below. The bar chart shows that the Home Workout App has the highest positive reviews, followed by the Google Fit



App and then the Samsung Health App. This shows that the Home Workout App is the most reliable of the three (3).

Fig 7:Total Sentiment Count of all Health Apps

5. CONCLUSION

This work has been able to validate the three (3) Health Fitness Applications considered using Users' Reviews. A total of Five thousand (5,000) reviews each was used for the Apps. Sentiment Analysis was used as an automated method of determining whether a usage-produced text conveys a positive view of the App. Analysis showed that the Home Workout Fitness Application had more positive reviews compared to the other two. It garnered a total of 99.9% Positive Reviews and 0.1% Negative Reviews. Users seem to be pleased with all the services being offered by this particular App.

Google Fit App came second in terms of validity with 37.4% Positive Reviews and 62.6% Negative Reviews. It can be said that it is moderately valid though not anywhere close to the Home Workout App.

Samsung Health Fitness App has the least Positive Reviews of 3.4% and most Negative reviews of 96.6%.

COMPETING INTERESTS DISCLAIMER:

Authors declare that no competing Interests exist. The research was not funded by anyone rather it was funded by personal efforts of the Authors.

REFERENCES

1. Carre˜no LVG, Winbladh K. "Analysis of user comments: an approach for software requirements evolution". In IEEE Proceedings of 35th International Conference on Software Engineering (ICSE). 2013:582- 591.

2. Ghosh A. "Google Fit Preview SDK now available". Google Developers Blog. 2014; Retrieved August 29, 2018.

3. Google Fit comes to iOS. TechCrunch. Retrieved April 25, 2019.

4. Google Fit is now on iOS. Google. April 24, 2019. Retrieved April 25, 2019.

5. Hollendoner M. "Introducing the new Google Fit". The Keyword. 2018; Retrieved August 21, 2018.

6. Jee YL. User Review Mining: An Approach for Software Requirements Revolution. International Journal of Advanced Smart Convegence 2020;9(4):124-131.

7. John PH. Smartphone Applications for Patients' Health and Fitness. Am J Med. 2016;(129)1:11-19.

8. Joseph RP, Durant NH, Benitez TJ et al. Internet-based physical activity interventions. Am J Lifestyle Med. 2014;8(1):42-68.

9. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. 2014;5(4):1093–1113.

10. Onan A, Serdar K, Hasan B. "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification". Expert Systems with Applications. 2016;62:1-16.

11. Phillips AC, Der G, Carroll D. Self-reported health, self-reported fitness, and all-cause mortality: Prospective Cohort Study. British Journal of Health Psychology. 2010;15(2):337-346.

12. Prasetyo BE, Putri DGP, Pamungkas EW. “Aspect Extraction using Informative Data from Mobile App Data Review”. International Journal of Computer Applications. 2017;173(9):28-32.

13. Sakitha AJ, Reshma RK, Sony V. User’s Perspective about Mobile Fitness Applications. International Journal of Recent Technology and Engineering. 2020;8(6)3368-3373.

14. Salvador A, Jeronimo G, Irena V, Moises G. The Intention to use Fitness and Physical Activity Apps: A Systematic Review. Sustainability. 2020;12:6641.

15. Samsung Introduces S Health Application for GSIII | SAMSUNG UK". Samsung UK. Retrieved 27 January 2018.

16. Seyff N., Florian G. and Neil M. ”Using Mobile RE Tools to Give End-Users Their Own Voice.” In 2010 18th IEEE International Requirements Engineering Conference. 2010:37-46.

17. Shupei Y, Wenjuan M, Shaheen K, Wei P. Keep Using My Health Apps: Discover Users’ Perception of Health and Fitness Apps with the UTAUT2 Model. Telemed J E Health. 2015;21(9):735-741.

18. Spittaels H, De Bourdeaudhuij I, Vandelanotte C. Evaluation of a website-delivered computer-tailored intervention for increasing physical activity in the general population. Prev Med. 2007;44(3):209-217.

19. The Verge, "Google Fit is getting redesigned with new health-tracking rings". Retrieved August 29, 2018.

20. Yuanchun L, Baoxiong J, Yao G, Xiangqun C. Mining User Reviews for Mobile App Comparisons. Proceedings of the ACM on Interactive, Mobile,Wearable and Ubiquitous Technologies (IMWUT). 2017.

21. Liu B, Lei Z. ”A survey of opinion mining and sentiment analysis.” In Mining text data. Springer, Boston, MA. 2012:415-463.