Toward Automated Categorization of Mobile Health and Fitness Applications

Qiang Xu
Dept. of Computing and Software
McMaster University
1280 Main St. W., Hamilton, ON
Canada
xuq22@mcmaster.ca

George Ibrahim
Dept. of Computing and Software
McMaster University
1280 Main St. W., Hamilton, ON
Canada
georgeibrahim.89@gmail.com

Rong Zheng
Dept. of Computing and Software
McMaster University
1280 Main St. W., Hamilton, ON
Canada
rzheng@mcmaster.ca

Norm Archer
DeGoote School of Business
McMaster University
1280 Main St. W., Hamilton, ON
Canada
archer@mcmaster.ca

ABSTRACT

In recent years, with the explosive adoption of smart phone devices, mobile health and fitness applications have been increasingly used by healthcare practitioners and the general public to manage electronic health records, chronic medical conditions, dietary references etc. Despite the rapid growth in the number of mobile and fitness applications on various platforms, very little work has been done to quantitatively and qualitatively assess these applications to guide users in the selection process. Automatic categorization of mobile health and fitness applications is the first step in this direction. In this paper, we report results from crawling 1,430 Android and 62,286 iOS apps in Nov. 2013. Among them, 1,399 apps were manually classified to one or multiple categories out of a total of 11 categories. Text mining tools were applied to the description section of the apps for keyword extraction, feature selection and automatic categorization. The classifiers we experimented with have comparable performance with Linear SVC achieving the highest precision, recall and f1 scores of 0.89, 0.79 and 0.88, respectively.

Categories and Subject Descriptors

D.3.3 [**Computing methodologies**]: Natural language processing, supervised learning; [**Human-centered computing**] Mobile computing

General Terms

Algorithms, Measurement

Keywords

Mobile applications, health and fitness, automated categorization

MobileHealth'14, August 11, 2014, Philadelphia, PA, USA.
Copyright © 2014 ACM 978-1-4503-2983-5/14/08 $15.00.
http://dx.doi.org/10.1145/2633651.2633658

1. INTRODUCTION

In recent years, smart phone adoption has increased rapidly, accounting for 58% of the US adult population by January 2014. Due to their advantages in packing communication, computing and sensing capabilities in one compact platform, smart phones have also gained popularity among healthcare professionals and the general public for health related applications. Many mobile medical and fitness apps are now available, transforming smart phones into tools to facilitate medical education, disease self-management, and clinical communication between healthcare providers and patients. According to Fox [1], half of all smart phone owners use their devices to get health information and one-fifth of smart phones owners have health apps. According to Laird [2], 247 million people downloaded a health app in 2012, compared to 124 million in 2011. This indicates that mobile health apps are growing rapidly in popularity and people have become aware of the benefits of such apps in managing their health.

According to AppBrain [3], there were 1,190,107 Android apps in the market, including 49,084 apps in the Medical and Health & Fitness categories as of November 21st 2013. No exact number for iOS apps is available but many websites claim that it exceeds 1 million. Despite the rapid growth in the number of mobile and fitness applications on various platforms, very little work has been done to quantitatively and qualitatively assess these applications in order to guide users in the selection process. Unlike other types of mobile apps, the quality of mobile health and fitness applications can have implications in user long-term health and wellbeing, and should therefore be subjected to a high degree of scrutiny. In fact, recognizing the rapid pace of innovation in mobile apps, and the potential benefits and risks to public health represented by these apps, in 2013 the US Food and Drug administration issued guidelines to clarify the subset of mobile apps (called *mobile medical* *applications*) to which the FDA intends to apply its authority [4].

Indeed, all the Apps Stores including Apple Apps Store and Google Play provide some degree of categorizations. Their taxonomies, however, are limited and too coarse to provide enough guidelines. Take Apple Apps Store as an example, it only has two health-related categorizations Health & Fitness and Medical. Considering their huge numbers, along with their simplistic taxonomy, there are increasingly difficulties not only for users to use but also for government to manage these health and fitness Apps. Therefore, coming up with a solution to provide refined level of categorizations is imperative. This study is the first of its kind in developing tools for automated categorization of mobile health and fitness applications on both Android and iOS platforms. A web-based crawler was developed to collect the description, user reviews and other meta-data information of 1,430 Android and 62,286 iOS apps in Nov. 2013. Among them, 1,399 apps were manually classified to one or multiple categories out of a total of 11 categories, based on domain-specific knowledge. The description of each app was then processed and keywords were extracted using the open source natural language processing (NLP) tools NLTK [5], WordNet [6] and Scikit-learn [7]. The term frequency–inverse document frequency (TF-IDF) and category labels were used to train classifiers via supervised learning. We experimented with Linear SVC [8], NearestCentroid [9], Naïve Bayes [10], and found that these three classifiers have comparable performance, but with Linear SVC achieving the highest precision, recall and f1 scores of 0.89, 0.79 and 0.88, respectively. Linear SVC was then applied to the remaining unlabeled apps for automated categorization. A comparison between the top keywords determined by TF-IDF and manual methods showed significant overlap, while at the same time indicating directions for improving NLP and machine-learning techniques.

To highlight some findings from the categorization results, we observe that:

- There are non-negligible percentages of foreign language mobile health and fitness apps on both Android (1%) and iOS (4%) platforms.
- Exercise and fitness apps are the most popular in terms of both the percentage of apps as well as the number of downloads. These are followed by healthy eating and weight loss apps, and reference apps.
- A number of apps belong to multiple categories that serve multiple purposes.

The remainder of the paper is organized as follows. In Section 2, we present web crawler and key statistics of the obtained datasets. The automated categorization system and algorithms are described in Section 3. In Section 4, the categorization results are presented along with discussion of limitations and key observations. The paper concludes in Section 5 with an ongoing research agenda.

2. DATA COLLECTION

A web crawler was developed that collects app information from the medical, health and fitness categories from Google Play and iTunes websites. To overcome the rate limits of Google Play, multiple crawlers were run in parallel to issue HTTP requests at a rate of 25 seconds per app. Since the start of the study, Google Play has made changes to its web interface so it only shows the top apps. In both Google Play and iTunes, apps were broken down into free and paid categories. The same app may appear in both categories, with added features for the paid categories. In Google Play, all user review information can be obtained. In contrast, only a limited number of user reviews are available inform the iTunes website (though more are available on App store on smart phones).

In addition to app descriptions and user reviews, both stores provide meta-data information regarding each app (see Table 1). The fields that differ between the two stores are highlighted in boldface.

Table 1 App Meta Data

Android	iOS
Developer/seller	Developer/seller
Date of update	Date of update
Ranking	Version
Total number of reviews	Total number of ratings
Size	Size
Current version	**Language**
Range of installs	Compatibility
Content rating	Custom ratings
Screen shots	**Customers also bought**
	Screen shots

As of November 24th 2013, data on 62,286 iOS apps and 1,430 Android apps data were collected from both the Health & Fitness and Medical categories using the developed tool. 57,996 of the iOS apps and 1,410 of the Android apps had a description field in the app webpage while the rest did not. A preliminary analysis of the iOS apps shows that there are 8,900 apps in common between the Health & Fitness and Medical categories, resulting in 49,096 unique iOS apps from both categories. There are 1,410 unique Android apps in the dataset. A comparison of the two sets, found 311 apps in common between iOS apps and the top Android apps for both Health & Fitness and Medical categories. The results show that about 96% of the iOS apps are available in English. Some are also available in different languages such as French, German, Spanish, Arabic, Chinese and others. However, about 4% of the iOS apps were only available in languages other than English. After reviewing the descriptions of the Android apps, it was found that about 99% of the apps were available in English while 1% of the apps were available in other languages. Figures 1 (a) – (d) show the range of downloads for Android apps.

Both Google Play and the App Store break down the mobile health apps into two broad categories of medical vs. health and fitness apps. Clearly, such a division is too coarse grained and is not very informative. After reviewing the descriptions the downloaded apps and consulting with domain experts, we came up with 11 sub-categories: *exercise and fitness, healthy eating and weight loss, reference, women's health, reminders and alerts, symptom checker, mental health, sexual health, pet health, personal health record/electronic health record, and disease monitoring.*

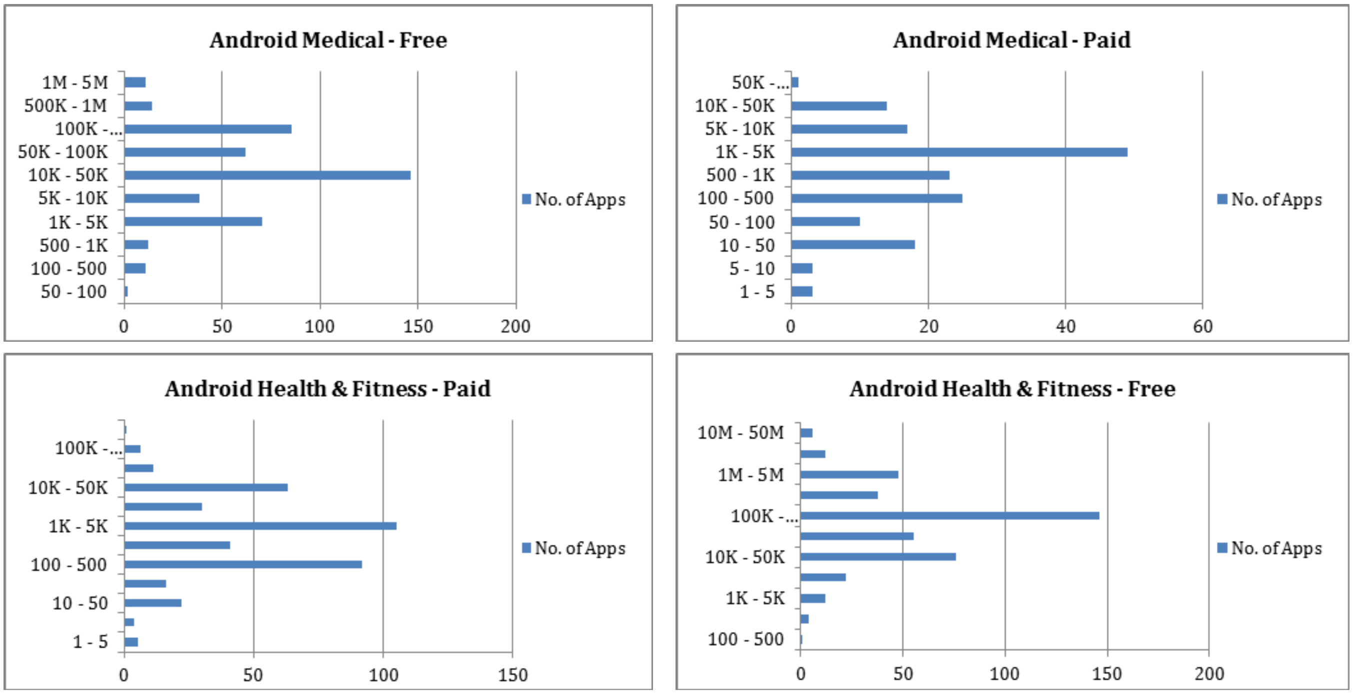

Figure 1 Installation history of Android Apps

3. AUTOMATED CATEGORIZATION METHODS

3.1 System architecture

Manually categorizing apps into different classes is tedious and time consuming. Thanks to the development of NLP (nature language processing) and machine learning techniques, it is practical to make use of text classification techniques to automatically categorize apps into different classes according to their descriptions. The data flow of the categorization system is summarized in Figure 2.

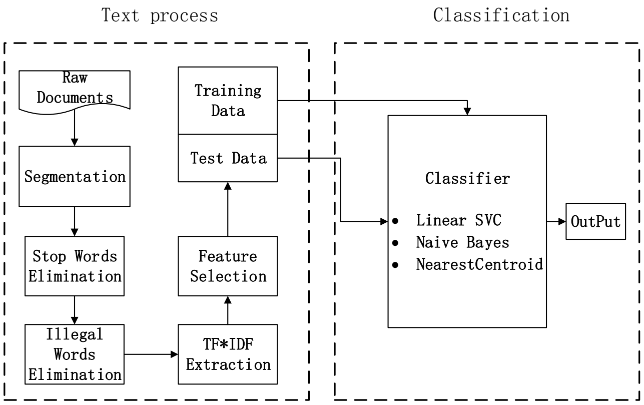

Figure 2 Data Flow of Automated Categorization System

There are two key stages: text processing and classification. Next is a brief description of the steps in each stage.

Text processing: The raw description of apps cannot be directly used for classification. They must be structured, and represented using a feature vector. To achieve this goal, several steps are required.

- Step1: *Text segmentation* (also known as, *text tokenization*). Since the raw text paragraphs are unstructured, we need to divide those paragraphs into meaningful units (words). For the English language, word boundaries are explicit and thus it is simple to perform text segmentation.
- Step2: *Stop word elimination*. Stop words are short function words such as 'the', 'a', 'which', 'who', 'am', 'is' and 'are'. Although in some particular situations, those words are meaningful, such as 'Dr. Who', the majority are useless and can be removed directly. Here, we remove all stop words.
- Step3: *Illegal English word elimination*. It is inevitable that there exist some illegal English words in text documents. Typos are common. Additionally, there are numbers and ordinals in app descriptions. In this step, we remove all numbers, any words whose size is less than 3, and finally, words that cannot be recognized as meaningful English words.

- Step4: *Word stemming*. This step extracts the root of each word. In English, a word may have different tenses or forms (plural vs singular). Although there are many complex models to process words in different tenses, stemming them directly is still the simplest and most effective. As an example, both 'exercises' and 'exercise' can be stemmed to 'exercis'.
- Step 5: *TF-IDF Extraction*. The previous steps extract meaningful features. But to structure the document, we still need to weight these features. The most straightforward way is to use the term-frequency. However, some very frequent but meaningless terms may overshadow the frequencies of rarer yet more meaningful terms. IDF*TF is widely used to tackle this problem. TF stands for term frequency, and IDF means inverse document frequency calculated by dividing the total number of documents by the number of documents containing a term, and then taking the logarithm of that quotient. A frequent term (e.g., health) that appears in most descriptions would thus receive a lower weight.
- Step 6: *Feature selection*. The set of text documents may contain thousands of different words. Not all of them are informative. Including all as part of the feature vector will lead to the curse of dimensionality. Therefore, feature selection needs to be applied. One common feature selection method in text mining is the Chi Square test [11], which is used to test whether the occurrence of a specific term and the occurrence of a specific class are independent. If they are dependent, then the term will be selected as a feature.

After pre-processing, each app description is represented by a feature vector of length 1000. The element of the feature vector is the IDF-TF of the respective term. We manually labeled (classified) 1399 apps and used them to train and validate the classifiers. We need to keep in mind that a sample size of 1,399 is not sufficient, with a feature vector of 1000, especially due to multiple categories.

Classification: Classification or categorization is one of the most typical supervised learning problems. We compared the performance of three classification algorithms. The first algorithm is linear SVC [8](support vector classifier) in which the support vector machine is used. SVC is effective in high dimensional spaces, and is suitable when the number of dimensions is greater than the number of samples. To implement multi-class classification, we used the 'one-against-one' approach [12]. The second classification algorithm chosen was NearestCentroid(NC) classifier [9]. When used for text classification with TF-IDF vectors, this classifier is also known as the Rocchio classifier. In this model, each class is presented by its centroid, with test samples classified to the class with the nearest centroid. The last algorithm used was the Naive Bayes(NB) classifier [10], based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features.

3.2 Implementation and performance

To implement the proposed classification system, we utilized existing open source NLP and machine learning tools. The entire system is Python based. We used NLTK [5] to segment the documents and stem the words. WordNet [6] was used to remove illegal English words. Scikit-learn [7] was used to calculate the TF-IDF weight and to train the classifiers.

The total number of text descriptions was 1399. The distribution of the apps across different categories was imbalanced as indicated in Table 2. For example, there were only 3 symptom checker apps compared to 354 exercise and fitness apps. The number of apps that belong to two or multiple categories is too low in the labeled data for training. As a result, we omitted these apps from the training set. Altogether, 80% of the labeled descriptions were used as the training set; the remainder were used as the test set.

Table 2 Size of Training and Testing Sets

Category	# of total samples	# of test samples
Exercise & fitness	354	72
Healthy eating & weight loss	317	65
Symptoms checker	12	3
PHR/EHR	12	4
References	287	59
Sexual health	15	4
Women's health	166	35
Reminders & alerts	58	13
Disease monitoring	57	13
Pet health	12	4
Mental health	100	21
Mental health + Exercise & Fitness	1	0
Healthy Eating & weight loss + References	1	0
References + Mental Health	1	0
Sexual Health + References	2	0
Healthy eating & weight loss + Exercise & Fitness	3	0
Women's Health + References	1	0

'+' means AND; the shaded rows are omitted in training the classifiers and the final evaluation.

Figure 3 compares the precision and recall of different classifiers for each category in the test data. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. The average f1 score, defined as $2\frac{precision \cdot recall}{precision+recall}$of linear SVC, NC and NB are respectively 0.88, 0.838 and 0.838. We conclude that linear SVC gives the best performance. This is likely to be due to its ability in handling imbalanced data.

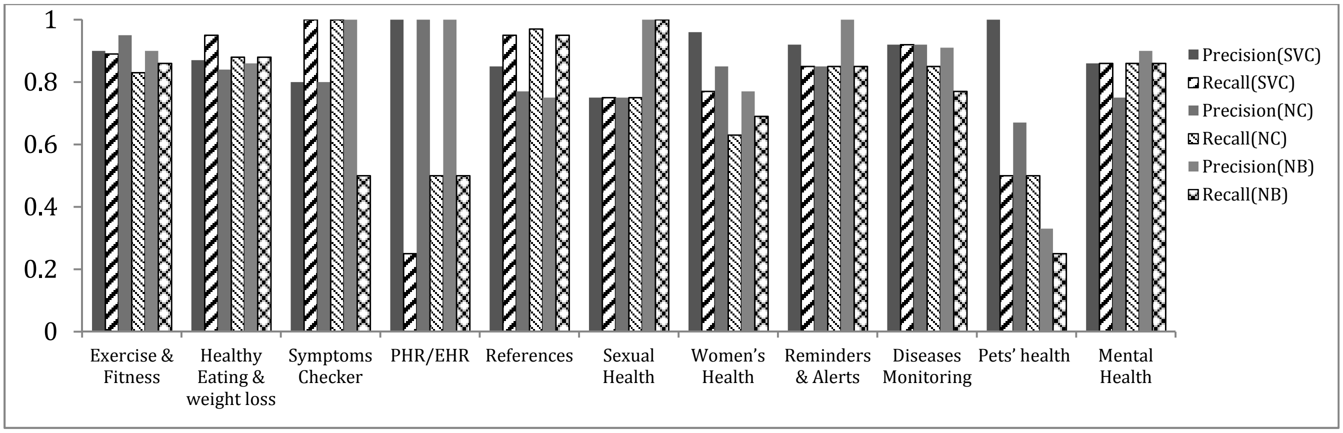

Figure 3 Precision and Recall of SVC, NC and NB Classifiers

4. CATEGORIZATION OF MOBILE HEALTH APPS

In this section, we apply linear SVC to the remaining apps and discuss the findings from the datasets.

4.1 Categorization results

Figures 4 and 5 show the breakdown of the iOS and Android apps in different categories. Clearly, for both platforms, the exercise and fitness category has the most number of apps followed by health eating and weight loss (iOS) or reference apps (Android). Reference apps provide different kinds of medical references and education tools, as well as drug information and interactions. One example of a medical reference app is "Human Anatomy Pro" that illustrates the anatomy of the human body and the different biological systems. It also provides external links to access more medical information on different websites. It is unclear why there is a higher percentage of reference apps in Android compared to iOS.

4.2 Top key words

An intuitive way for humans to categorize mobile health and fitness apps is to use key words. One can come up with a list of keywords for different categories. If one or more of the keywords are found in the text description of an app, it is assigned to the respective category. Table 3 compares the list of keywords selected by a human and the top-10 (by TF*IDF) determined by the NLP algorithms. We observe that many of the manually selected keywords also appear among the top-10 machine generated keywords. We also found some keywords identified by the NLP algorithms that were not previously recognized by humans. For example, in the sexual health category, "yoga" is among the top-10 keywords. There appears to be some relation between yoga practice and sexual health. On the other hand, we also discovered some limitation of our machine learning based approach. For example, BMI (body mass index) is clearly relevant to healthy eating and weight loss but was excluded from the keywords since it is not a legitimate English word. This can be addressed by including medical terminologies and/or phrases in feature extraction.

Table 3 Comparison Between Manually Selected Keywords and Machine Generated Keywords

Category	Manually Selected keyword	Top-10 keyword (TF*IDF)
Exercise & Fitness	Exercise, Fit/Fitness, Workout, Run, Body + Build + muscle, Gym	Workout, Exercise, Fit, Train, Use, Time, Body, Run, Program, Weight
Healthy Eating & weight loss	Food + Diet, Calorie + Count + fat, Calorie+ Calculate + metabolic, Weight + loss/lose + food, Recipe, BMI + Calculate, Meal + plan, Nutrition	Food, Recipe, Diet, Weight, Eat, Calorie, Gluten, Day, Meal, Restaurant
Symptoms Checker	Symptom + check, Symptom + guide, Symptom + manage	Symptom, Medic, Care, Doctor, Use, Provide, Advice, Checker, Help, Inform
PHR/EHR	PHR, EHR, EMR, Electronic + Health + record, Electronic + Patient + record , Electronic + Medical + record, Personal + Health + record	Health, Medic, Record, Track, Inform, Patient, History, Use, Data, Vital
Reference	Dictionary, Handbook, Education, Atlas, Reference, Guide, Medical + term, Drug + information/interaction	Drug, Medic, Dictionary, Use, Information, Atlas, Handbook, Miscellany, Application, Animated
Sexual Health	Sexual + guide/tip/info	Sex, Sexual, Yoga, Std, Life, Use, Orgasm, Application, Animated, Enhance
Women's Health	Woman/Women, Pregnancy/Pregnant, Period + track, Menstrual, Ovulation	Pregnant, Period Women, Day, Baby, Ovulation, Health, Fertile, Use, Cycle
Reminders & Alerts	Reminder, Alarm , Appointment + remind, Pill + remind	Alarm, Sound, Sleep, Noise, Clock, Pill, Timer, Wake, Time, Use
Diseases Monitoring	Monitor /track/manage + Blood pressure/BP, Monitor /track/manage + Diabetes/glucose, Monitor /track/manage + heart rate/cardio, Monitor /track/manage + chronic disease	Blood, Pressure, Diabetes, Glucose, Track, Measure, Sugar, Monitor, Read, Weight
Pets' health	Pet , Dog/cat, Veterinary	Dog, Pet, Animal, Drug, Veterinary, Shaw, Wildlife,

		Emergency, Use, Know
Mental Health	Mental + Health, Anxiety, Stress + Relief/manage, Depression, Sleep, Relax/Relaxation + music	Anxiety, Sleep, Stress, Use, Mental, Disorder, Relax, Depress, Medication, Health

4.3 Limitations

This study provides the first step in the automated categorization of mobile health and fitness apps. There are a number of limitations in the current study that will be addressed in our future work. Specifically, in terms of the datasets, we only obtained a small set of top Android apps due to the restriction of the Google Play website. These apps may not be a good representation of all categories. For both iOS and Android apps, only apps in English were considered. In this study, only text description of the apps was utilized. It is expected that other sources of information such as user reviews, snapshots of the apps, or even the executable files of the apps would provide more insights. In terms of methodologies, the 11 categories may not be exhaustive. Other ways of categorization may be possible such as those based on types of health conditions. Finally, as discussed earlier, machine learning and NLP techniques themselves can be improved to better handle medical terms, imbalanced datasets, and overlapping categories.

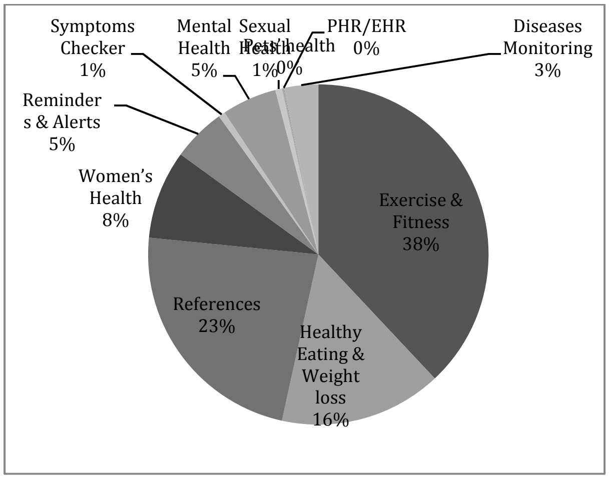

Figure 4 Categorization of Android Apps

5. CONCLUSION

In this paper, we presented a framework for automated categorization of mobile health and fitness applications and demonstrated its effectiveness by crawling iOS and Android apps from Apple App Store and Google Play. The categorization provides a quantitative understanding of available apps, which constitutes an important step in the qualitative assessment of these apps for enhancing health and wellbeing of the general population. In addition to addressing the limitations outlined in Section 4, as future work, we will build tools to evaluate the performance of various apps and provide recommendations for different categories.

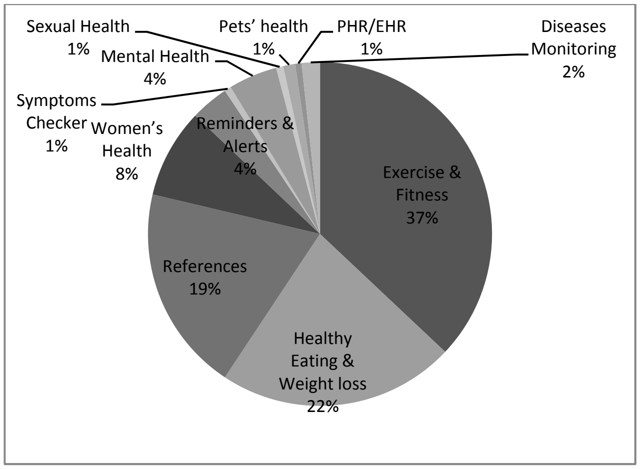

Figure 5 Categorization of iOS Apps

6. ACKNOWLEDGMENTS

This work is in part supported by NSERC Discovery Grant #2012R00435 and Ontario Research Fund - Research Excellence.

7. REFERENCES

[1] Fox, S., Duggan, M., 2012. Mobile Health 2012: Half of Smartphone owners use their devices to get health information and one-fifth of Smartphone owners have health apps, Pew Internet & American Life Project, California Healthcare Foundation.

[2] Laird, S. 2012. How Smartphones are changing healthcare, http://mashable.com/2012/09/26/smartphones-health-care-infographic/

[3] http://www.appbrain.com/

[4] U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Center for Biologics Evaluation and Research, 2013. Mobile Medical Applications: Guidance for Industry and Food and Drug Administration Staff.

[5] http://www.nltk.org/

[6] http://wordnet.princeton.edu/

[7] http://scikit-learn.org/stable/

[8] Gunn, S. R. 1998. Support vector machines for classification and regression. ISIS technical report, 14.

[9] Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America, 99(10), 6567-6572. The National Academy of Sciences.

[10] H. Zhang, 2004. The optimality of Naive Bayes. Proc. FLAIRS.

[11] Greenwood, P.E.; Nikulin, M.S. , 1996. A guide to chi-squared testing. New York: Wiley. ISBN 0-471-55779-X.

[12] Knerr, S., Personnaz, L., and Dreyfus, G., 1990, Single-layer learning revisited: A stepwise procedure for building and training neural network. Neurocomputing: Algorithms, Architectures and Applications, NATO ASI, Berlin: Springer-Verlag.