Spring 2017

Trading Data for Discounts: An Exploration of Unstructured Data Through Machine Learning in Wearable Technology

Kaleigh Alexis Zukowski
University of New Hampshire, Durham, kau549@wildcats.unh.edu

Trading Data for Discounts: An Exploration of Unstructured Data Through Machine Learning in Wearable Technology

Kaleigh Zukowski

Department of Decision Sciences

University of New Hampshire

Thesis Advisor: Inchan Kim

Table of Contents

Abstract

The development of computing sensor devices with the capability of tracking an individual's activity changed the way we live and move. The data collected and generated from wearable technology provides implications to the user for leading a healthy, more active lifestyle; however, the potential data uses extend beyond the user. Significant opportunity exists in the insurance industry as it relates to discounting premiums. The purpose of this research was to provide insight as to whether insurance companies should consider offering discount on premiums for policyholders who use wearable technology to track their personal fitness by identifying and suggesting potential groups of consumers to target these discounts toward. Using the platform R, researchers collected and analyzed tweets about four leading wearable technology companies including Fitbit, Jawbone, Misfit, and Withings. Both unsupervised and supervised learning techniques were pursued during the study in the form of topic modeling and artificial intelligence. Through detailed analysis, researchers determined that companies may want to consider reducing premiums for wearable technology users who use the devices for weight loss, as it would benefit both policyholders and insurance companies.

Introduction

Measurement is timeless. From qualitative to quantitative, the concept maintains a critical place in our culture. Distance, weight, time, and progress are among the valuable metrics calculated. Maintaining such records provide insight into ourselves and the world around us, allowing for an understanding of normalcy and deviancy. Gathering data measurements of an individual was introduced through healthcare and the inception of the weight scale. Throughout time, physical quantification transitioned from a scale at the doctor's office to a household device, and advanced to simply wearing a wristband containing computing capabilities. Centered around the individual's desire to achieve self-improvement, wearable technology, although more modern than the weight scale, encompasses the same foundation of gaining awareness of the individual body (Crawford, Lingel, & Karppi, 2015).

Wearable technology provides insight into the individual's level of activity by tracking fitness metrics including steps taken, distance travel, calories burned, and minutes of activity. Immediate access to real time information attributes to the adoption of the technology, as users find that it motivates them to lead a healthier lifestyle and participate in fitness activity. While wearable technology is often designed with a focus on the individual, the data generated by the devices have the potential to be used by a wider range of entities. Significant opportunity exists for wearable technology to be integrated into the insurance industry. Access to actual policyholder data could justify discounts and the reduction of insurance premiums for those individuals taking action to improve their health and quality of life.

I. Research Justification

This research seeks to identify an answer to the following question:

"Should insurance companies provide discounts to policyholders who use health tracking technology in their daily lives? If so, which segments of policyholders should insurance companies target?"

Research was designed to explore the area of health-related data, gathered using wearable technology, and its impact on insurance premiums. The study aimed to provide an answer based on the identification of potential consumer groups in which insurance companies could target their discounts toward. Throughout the duration of the research, two objectives were pursued: unsupervised learning through topic modeling and supervised learning through artificial intelligence. With the use of unstructured data involving four leading companies within the wearable technology space, researchers gained insights as to the context of the proposed research question.

II. Company Backgrounds

The research utilized consumer data associated with four prominent companies within wearable technology, Fitbit, Jawbone, Misfit, and Withings. Although all four companies exist in the same market and compete, each company is characterized by its own identity and product offerings. Stemming from unique backgrounds, the companies are well established in their values and have passionate consumers supporting their operations.

Fitbit strives to encourage users to create a lifestyle centered around activity and encourage them along the path toward health and fitness. With devices and experiences that can uninterruptedly integrate into everyday life, Fitbit users can achieve their goals. The company's co-founders launched their adventure in 2007 with the realization that sensors and wireless technology could enhance the fitness and health experience. The development of Fitbit transformed human motion and lifestyles (Fitbit, 2017).

Jawbone sets the pace in wearables and consumer technology through the development of products and software platforms. The company first developed a wireless headset and accompanying software, which was followed by the creation of the Jambox suite, a series of Bluetooth capable speakers. Before long, Jawbone introduced UP, its lifestyle tracking system, expanding its reach to the wearable technology sector in 2011. Since its inception, UP has advanced and developed, consistently supporting users through the development of personalized awareness of oneself with data on sleeping, eating, and moving (Jawbone® | About The Company, 2017).

Created in 2011, Misfit Wearables specializes in wearable technology with a focus on design. Its original device, Shine, sparked a passionate following, allowing Misfit to continue its journey of innovation. The company enhanced its wearable technology, incorporating advanced features to meet the needs of its users, and evolved beyond fitness trackers. Misfit has since developed an exclusive sleep monitoring device and a smart lightbulb (Misfit: Fitness Trackers & Wearable Technology, 2017).

Withings entered the technology industry with the first Wi-Fi scale in 2009 and has been a leader among creators of smart products and applications. Incorporating design and innovation into its technology, Withings aims to create products for any lifestyle. The company encourages users to become healthier and more equipped for maintaining a productive and active life. Withings connects the individual to the health revolution (Withings, 2017).

III. Organization of Thesis

The remainder of the thesis is organized according to the following functional areas: literature review, methodology, results, discussion, and conclusion. The literature review examines the current state of the wearable technology environment and the insurance industry. It provides insight into extant interactions between the two entities as they relate to the designed research. Following the review, the thesis explains in detail the research methodology executed and the results generated from those tactics. An analysis of the results is provided in the discussion section along with a discussion of the research limitations. In the final section, conclusions are drawn from the discussed results. Relevant supplementary materials are included in the appendices.

Literature Review

As technology advances, trends fluctuate and users adopt the latest and greatest innovation. The recent movement toward the interest in monitoring daily activity triggered the integration of wearable technology into everyday life. Wearable technology consists of two components, a small hardware device and an application. Its capabilities focus on activity tracking metrics including steps taken, distance traveled, and calories burned; additional features such as sleep monitoring and heart rate tracking are included in some advanced devices. The wearable instruments are state-of-the-art computers, on a much smaller scale, that are worn on the body as clips, bracelets, wristbands, or smartwatches. The computing apparatus contains the capacity for electronic monitoring and syncs wirelessly to its application stored on a smartphone or computer, allowing for long-term data management (Kaewkannate & Kim, 2016).

The ability to retrieve personal data in real time contributed to the significant acceptance of wearable technology. Automatic access to monitoring and tracking routine tasks in a user's day replaces the manual calculations of workouts and allow for the possibility to enhance life through health, food, and fitness. The trackers serve as motivators in personal healthcare, encouraging users to be more active and strive to maintain healthy lifestyles. Users can set goals and use the devices as personal assistants in achieving their ambitions. The variety of users of wearable technology contribute to the need for a vast array of product offerings; each device is tailored to the unique personal needs of different consumers (Kaewkannate & Kim, 2016).

The United States is home to millions of fitness tracking fanatics who use the devices for step counting and motivation. The individual is not the only population adopting fitness tracking; employers have developed programs based around the accessory to reward their employees for achieving the goals they set. The technology is also gaining speed within the insurance industry, as life and health insurers are beginning to offer discounts on premiums for those who incorporate the technology into their lives (Pitsker, 2016).

One employer utilizing wearable technology in its corporate wellness program is startup company, Appirio. The company writes cloud applications for organizations, specifically in the healthcare industry. Its corporate wellness program, CloudFit, distributed Fitbits to 400 employees that were purchased by Appirio with funding from their insurance company. The program also gave employees access to sessions with personal trainers using Google Hangouts. One quarter of the participants voluntarily shared their data with the company, but decided which metrics to share. The wearable technology users gained access to a social networking group developed for challenging and motivating their peers (Bort, 2014).

Being in the business of writing applications, Appirio created a cloud-based app to aggregate the data generated by their employees using Fitbits as a part of their program. Using this information, Appirio convinced its insurance company to lower its price. With the CloudFit program in place, Appirio received a five percent discount on its rates. This reduction cut company costs by $280,000. Appirio’s employee engagement followed in the footsteps of BP America, who developed a similar program the year prior. BP distributed Fitbits to 24,000 participants, which included involvement from employees, spouses, and retirees as a part of their corporate wellness strategy (Bort, 2014).

The insurance industry is characterized by rising costs, complicated benefit structures, and the denial of coverage, which contribute to the characterization of insurance providers as universal enemies. Insured consumers often associate interactions with their cable and internet providers as more positive than interactions with their insurance companies. The aversion to insurance companies spreads beyond policyholders within the health care sector; providers and physicians share a mutual dislike toward the entities due to the considerable negotiation required for the pricing of services and the administrative documentation necessary for each patient and appointment. With significant question surrounding the insurance industry, the opportunity exists to eliminate the distrust and improve customer perceptions and relations (Soderland, 2015).

Health insurers are situated in a position of responsibility and knowledge; they manage much of the health care expenses, defining their power in the market, and they possess insight as to the customer segmentation of their products. Such a situation provides the insurers the occasion to relieve the existing animosity through the development of innovative products, services, and partnerships centered around the needs of consumers. Specific consumer needs may include the focus on value, low-cost, personalization and convenience (Soderland, 2015).

Although many possible avenues exist, one specific idea insurance providers may choose to pursue is to incentivize healthy behavior change. A patient's lifestyle and personal behaviors determine fifty percent of the patient's health factors; an individual's health condition may be improved if he or she lives a life centered around health conscientiousness. The utilization of behavioral science and wearable technology in this space presents the opportunity for insurance providers to improve relationships with policyholders, through establishing partnerships and providing incentives, and increase cost savings, by encouraging healthy living among consumers. Together, these factors may improve the current negative perception associated with insurance providers (Soderland, 2015). While the idea of integrating insurance and wearable technology beyond the employee sector to individual policyholders remains relatively novel, some companies have increased their efforts and developed initiatives that incorporate the two.

John Hancock illustrated this idea; as a leading provider of insurance and financial service products, the firm recently partnered with Vitality, a company that consolidates benefits programs and wellness products. The partnership included the distribution of Fitbit devices to new policyholders to maintain healthy habits, thus reducing their risk of illness and death. The terms of the new policyholder's insurance agreement required the policyholder to participate in an online health review published by Vitality, which determines the individual's overall health and adjusts according to the person's health habits. Further insights into health habits may be gathered using the Fitbit device. Those consumers who purchase insurance products corresponding with the Vitality partnership are eligible to receive up to a fifteen percent discount on their insurance premiums as the policyholder improves their health. Designed as a smart form of life insurance, John Hancock management approached this initiative similar to an auto insurance policy based off of usage (John Hancock's Plan: Use Fitbit Wearables to Adjust Life Premiums, 2015). This initiative demonstrates John Hancock's interest in improving consumer relations and promoting positive partnerships.

With the evolution of technology comes frequent functional advancements. Initially developed as fitness trackers, wearable technology now encompasses the capacity of a more accessible smartphone. Reminders, notifications, and access to global news on the wrist eliminates the need to reach into a pocket for a smartphone. As functionality increases, so does the rate of adoption; in today's world, wearable technology is ubiquitous (de Waal, 2016).

Original wearable technology, as fitness trackers, had the potential for application to the insurance industry. Now that the devices are enhanced, however, the opportunities have increased. The data gathered by wearables can impact the underwriting and claims process and improve relations between companies and their policyholders. Providing insurance companies access to wearable technology data allows for more accurate evaluation of each policyholder, resulting in the creation of more representative quotes and the reduction of risk. The data collected from the devices can be used to confirm the validity of claims and accelerate the claims handling process. Claim arbitrators would be able to evaluate a claim in real time and make crucial decisions regarding action plans, medication, and rehabilitation instantly. The application of wearable technology to insurance also exists beyond the realm of fitness tracking. Advanced features, such as GPS capabilities, traffic alerts, and weather reports, could be used to prevent accidents and injuries by keeping policyholders safe (de Waal, 2016).

Wearable technology is a new standard in society, and it should be integrated into current practice. Taking advantage of the technology and analyzing its capabilities will allow companies to situate themselves in such a way to use the gadget in a safe and strategic manner. Using wearables to promote safe and healthy lifestyles among policyholders, while conceivably reducing premiums and payouts will benefit insurance companies and policyholders, thus improving relations and perceptions of the insurance industry (de Waal, 2016).

Methodology

The research conducted revolved around the platform R, a software environment and programming language used for statistical computing and graphics (What is R?, n.d.). The tool was used throughout the duration of the research to provide the basis of the analysis and discussion. The study was designed in three distinct phases, data collection, data preparation, and data analysis. The data analysis phase could be further broken down into unsupervised and supervised learning techniques.

I. Data Collection

Data collection was the initial task for the research performed. The data analyzed in this study stemmed from Twitter posts, known as tweets, about four leading companies within the wearable technology space including Fitbit, Jawbone, Misfit, and Withings. Each of the tweets were posted by product users and contained the hashtag followed by the company name, #Fitbit, #Jawbone, #Misfit, #Withings. The data was collected during a period of ten days in February of 2017 due to the capabilities of R.

Prior to gathering data, two setup efforts were executed: Twitter application creation, and R session preparation. It was necessary to create an application through the Twitter developer platform, which generated the personalized authorization tokens required to access the Twitter data. Once created, the tokens were included in the R code with the function **setup_twitter_oauth(***tokens***)**, allowing for the collection of tweets meeting some specific criteria. In addition to the creation of the Twitter developer application, researchers installed the required R packages that support tweet collection. Each R package is inherent to the platform and contains data, functions, and code for accomplishing a given objective; the packages installed included *"devtools"*, *"httr"*, and *"twitteR"* with the function **install.packages()**. Following the download and installation, a task that needs to be performed only once, the packages were loaded into the current R session with the **library(***package***)** function. Although packages require one time installation, each package must be loaded into every R session in which the package is required. With the Twitter authorization tokens completed and the R session prepared, data collection began.

Data collection involved three R functions, **searchTwitter()**, **twListToDF()**, and **write.csv()**. The **searchTwitter(***searchString, n, language, date restriction***)** function initiates a search of Twitter based on the search string provided by the user. In this case, the search string implemented was "*#fitbit*", "*#jawbone*", "*#misfit*", and "*#withings*" because research required tweets posted by product users of these four health tracking technology companies. After specifying the search string, we set the maximum number of tweets to be returned by the query, represented by n; we selected a large number to collect all possible tweets. The language was specified as English, and the date restriction was set as a date that occurred more than ten days previously. The **searchTwitter()** function is limited to collecting tweets posted within a ten day period from the date the function is run. Next, **twListToDF()** was run to generate a list of the tweets that met the search string defined. Following the list generation, we used **write.csv(***list, file location***)** to create an Excel spreadsheet for the tweets collected. The R code used to gather the tweets is provided in Appendix A, Section 1.

The code displayed in A.1 was run four different times to collect the tweets for each of the four companies. The only changes required to the code were in the **searchTwitter()** function, in the search string specifying the company hashtag, and the file name of the **write.csv()**. For organizational purposes, each Excel spreadsheet was titled with the company name. The gathering process resulted in the creation of four separate Excel spreadsheets that contained a total of 20,841 tweets, between all four companies, for analysis purposes. The tweets collected about Fitbit generated over 20,000 tweets, while Jawbone, Misfit, and Withings generated approximately 300 tweets each. The tweets collected were in the form of raw tweets as represented by a sample of Fitbit tweets in Figure 1 below.

Figure 1: Sample of Fitbit Raw Tweets

	A
1	text
2	Love My Watch <U+231A><U+FE0F>Username: msolisrazo@yahoo.com #fitbitcharge2 #fitbit #fitnessmotivation… https://t.co/R6rKv439vp
3	I climbed 4,000 floors with my #Fitbit and earned the 747 badge. OOOOOOHHHHHH SNAP. https://t.co/a3ZIfs3Jcq
4	No wonder my feet knack today! #fitbit https://t.co/82uGm2bHUf
5	<U+231A>#Fitbit #Watches #FitbitSurge #Fitness Activity Tracker Superwatch with Heart Rate Monitor (Small)… https://t.co/fc3SSw0EHB
6	Today = 547 consecutive days of #10000steps a day
7	I covered 26 miles with my #Fitbit and earned the Marathon badge. #killingit https://t.co/NGHpZRSZbt
8	I covered 70 miles with my #Fitbit and earned the Penguin March badge. https://t.co/rRA41osxXX
9	RT @Sunglassjunkies: Choose any #sunglassjunkie sunglasses RT them & tag us @sunglassjunkies for your chance to #WIN a #Fitbit Charge 2 htt…
10	#Super #Watches #FitbitSurge #Fitness Superwatch, Black, Large + #FitbitSurge #Fitness Superwatch… https://t.co/Mi304cj4SS
11	#Best #Watches BRAND NEW!!! #Fitbit #Surge SmartWatch Activity Tracker Heart Rate #GPS Black Large… https://t.co/Ob923LIwpR
12	#Health #Solution Burn Belly Fat Fast With These Pro Tips https://t.co/mZMCvDIU4h #Cavaliers #FitBit
13	History Sync for Fitbit to Apple Health #fitbit #gym #fitfam #workout #fitness #fitspo https://t.co/ki5D3SvDAd https://t.co/ZkFAvQTQRd
14	I need to charge my #Fitbit so I'm getting in as many steps as I can. And I'll be live tweeting the #Oscars
15	#GetFit #Fitbit What You Must Include In Your Weight Loss Plan https://t.co/QIQVlxXdJf #Dolphins #Workout
16	#Motivation #Fitbit How To Stay On Track With Your Fitness Goals? https://t.co/jqysUxPWr4 #Dodgers #Sporting
17	#Motivation #GetFit Burn Belly Fat Fast With These Pro Tips: Burn Belly Fat Fast With These Pro… https://t.co/zvfd5QZvyi #NickiMinaj #Fitbit
18	#WeightLoss #Fit How To Stay On Track With Your #Fitness Goals? https://t.co/EkQDz6dPGj #Celebs #Fitbit
19	#WeightLoss #Program 4 Weight Loss Side Effects And How To Deal With Them https://t.co/4m2GuqhxNL #Music #Fitbit
20	#FatLoss #Fit 4 Weight Loss Side Effects And How To Deal With Them https://t.co/0bmpNoUx0d #Heat #Fitbit
21	#Fitbit #GetFit 3 Different Ways to Lose Weight https://t.co/bHVHop3meu #KateMiddleton #Diet
22	#WeightLoss #Program Burn Belly Fat Fast With These Pro Tips https://t.co/VYGIlUOsrD #Style #Fitbit
23	#Motivation #Fitbit Burn Belly Fat Fast With These Pro Tips https://t.co/Co962l37lG #MeganFox #Workout
24	when all your co-workers are taking the #fitbit competition a little too seriously and you just can't take it anymore <ed><U+00A0><U+00BD><ed><U+00B9><U+008D><ed><U+00A0><U+00BC><ed><U+00BF><U+00BB> #stressedaf
25	Want to see what I think of my @FitbitUK? Head over to my blog for more info! https://t.co/ohCXoJj3Mj #fitness #fitbit #lbloggers #bbloggers
26	#Super #Watches #FitbitSurge #Fitness Super Watch #GPS, Heart Rate Monitor Black SMALL - Sealed!… https://t.co/zfQngCfCOQ

II. Data Preparation

Data preparation was executed by implementing the data preprocessing technique to ready the datasets for analysis. Data preprocessing is a method implemented in data mining that involves converting raw data into an understandable format, which in this case took the form of comprehensible text.

To prepare R for data preprocessing, researchers installed the packages "*tm*", "*SnowballC*" and "*slam*" with the function **install.packages()** and loaded each package into the current session with the **library()** function. We then used the function **read.csv()**, which specified the file R must read in order to run the given functions. Within the **read.csv()** function, we provided the location of the text files generated during the data collection phase; R read the Excel files containing the Fitbit, Jawbone, Misfit, and Withings tweets in the raw text form.

Next, we created the corpus, which represents a collection of text documents. For this research, the corpus is specified as the Excel spreadsheet identified by the **read.csv()** function. The corpus function, **Corpus(VectorSource(***file$column***)** requests the name of the text file and the column containing the relevant text. Once the corpus is created, the data is prepared to sustain the data preprocessing transformations.

By transforming the data, we applied multiple functions including **function()**, and **tm_map()**. Each of the functions transformed the raw tweet data into understandable text for analysis. Some of the transformations included the removal of unnecessary characters, such as punctuation, Twitter account names, numbers, URLs, and white space. We also removed stop words, or words that do not hold significance and are unnecessary for analytical purposes, including words such as "a," "and," "I," and "or." The complete list of transformations performed are included in the R code displayed in Appendix A, Section 2.

As with data collection, data preprocessing was applied to all four companies using the four spreadsheets developed in data collection. The code changes required involved the change in the file defined by **read.csv()** and the file and column specified by **Corpus(VectorSource())**. Once data preprocessing was complete and the tweets took the form of understandable text, they were prepared for further analysis through unsupervised and supervised learning techniques.

III. Data Analysis

A. Unsupervised Learning – Topic Modeling

The first aspect of data analysis involved unsupervised learning, a machine learning task that finds the structure or patterns between inputs assigned by the user. The unsupervised learning technique implemented in this study was topic modeling, a type of text mining and statistical model for revealing abstract concepts within a collection of documents. The R file created for data preprocessing served as the starting point for topic modeling. Since the package "*slam*" was previously installed during preprocessing, we loaded it in the current session with **library()**. Following package loading, we created a document term matrix with the function

DocumentTermMatrix() using the previously identified corpus. A document term matrix is a mathematical matrix that describes the frequency of terms that occur in a given corpus, or collection of documents. After transforming the corpus into a document term matrix, we ran a summary of the term-frequency (tf) scores assigned by the matrix. We then applied term frequency-inverse document frequency (tf-idf), a statistical device for information retrieval, to the document term matrix to eliminate frequent words that lack significant meaning, such as the term Fitbit found in the Fitbit tweets. A summary of the tf-idf scores were generated after applying the device.

The results of the tf and tf-idf scores were used to determine the term frequency values that would be of value to the study. In this case, it was determined that the terms associated with a term frequency between 0 and 0.10 would be evaluated. This threshold was established by evaluating the outputs R generated when running the topic modeling functions and the conclusion was made that the most valuable topics were generated when the 0 and 0.10 threshold was implemented.

Once the document term matrices were created and the threshold determined, the "*topicmodels*" package was installed and loaded into the current R session. In the topic modeling process, two statistical algorithms, VEM and Gibbs, were applied to each collection of tweets to reveal the abstract underlying topics within the data. The code developed for the topic modeling process is displayed in Appendix A, Section 3. Topic modeling was performed for each of the four companies and the output results generated were four tables of topics for each company. After developing the topic output values for each algorithm, researchers used R to assign a topic number to each tweet in the original document and generated Excel spreadsheets containing the information.

For each algorithm, two sets of topics were generated for each company, the tf terms and the tf-idf terms. For each company, four output tables were created, VEM tf, VEM tf-idf, Gibbs tf, and Gibbs tf-idf. In total, sixteen tables of topics were analyzed to make conclusions. Each list of outputs contained five different topics, but the number of terms presented in each topic vary among companies in order to eliminate words that did not have significant meaning, such as hashtags that were frequently used for a given campaign and provide no insight into the ways in which product users integrate wearable technology into their daily life. In addition to the creation of topic tables, we used R to create word clouds for the data. Each company has five associated word clouds, one representing the corpus which contains all terms in the document, two for the VEM algorithm, and two for the Gibbs algorithm. Of the two word clouds generated for each of the algorithms, one represents tf, while the other represents tf-idf. All outputs are displayed in the results section below.

B. Supervised Learning – Artificial Intelligence

The second component to data analysis dealt with supervised learning, the machine learning task of inferring a pattern from training data identified by the user. To initiate this process, researchers generated a random sample of two hundred Fitbit tweets and manually coded them into one of two categories, relating to fitness or not relating to fitness as seen in Figure 2.

Figure 2: Fitbit Random Sample

	A	B
1	Fitness	Text
2	TRUE	February 25, 2017 #Fitbit activity: 11745 steps taken, 8.77 kilometers walked/ran, and 2834 calories burned.
3	FALSE	#Motivation #Fitbit The Slim Person You Seek Is Inside You https://t.co/tB6XtOBNbR https://t.co/ylvsFBPcW3 #Dodgers... https://t.co/Jo1Yrdt3tX
4	FALSE	RT @bcSharonZ: #Fitbit is the new #Tamagotchis #narcissism https://t.co/mbGGhVlnE5
5	FALSE	Aha. That´s where all that email spam is coming from. #Fitbit! https://t.co/vftGNAHjFz
6	FALSE	#FatLoss #Fit Best Way To Burn Calories And Lose Weight – Cardio Versus Weight Training https://t.co/lFSUJVxPro #Heat #Fitbit
7	FALSE	Having a new toy to play with is fun mostly when it is #Fitbit
8	FALSE	#Top #Fitbit #Surge, Lightly Used, LARGE BLACK, Cable and Dongle Included https://t.co/719GjL0IyW #Sports #Deal https://t.co/vHea7E48uu
9	FALSE	#WeightLoss #Fit Being Fat Is Not Your Fault? https://t.co/yyOy1Mhe2N #Celebs #Fitbit
10	TRUE	Check out my trophy for getting first place in the Workweek Hustle challenge! #Fitbit https://t.co/ZQzB3UmKBy
11	TRUE	Never quit <ed><U+00A0><U+00BD><ed><U+00B2><U+00AA> #fitness #fitfam #fitspo #fitnessmotivation #fitbit #fitlife #fit #fitspiration... https://t.co/hSPlqO6z1t
12	FALSE	Cleaning out the van on this super nice day <ed><U+00A0><U+00BC><ed><U+00BC><U+009E><ed><U+00A0><U+00BD><ed><U+00B2><U+00AA> #rackingupmysteps #fitbit https://t.co/1Q1ZN5cGq8
13	FALSE	#Motivation #GetFit 7 Changes You Need To Make To Lose Stubborn Belly Fat... https://t.co/NZB9V7h4Se #NickiMinaj... https://t.co/YXqbalClls
14	FALSE	#Fitbit #GetFit 5 Tips On How A Woman Can Lose Weight Fast https://t.co/dumBLvt3P3 https://t.co/LsTTJOoSTi... https://t.co/6mf2fBIG1i
15	FALSE	History Sync for Fitbit to Apple Health #fitbit #fitfam #gym #apple #workout https://t.co/ki5D3SvDAd https://t.co/1wC7n76RyB
16	FALSE	History Sync for Fitbit to Apple Health #fitbit #fitness #fitspo #fitfam #iphone https://t.co/ki5D3SvDAd https://t.co/4WlAArzUYD
17	FALSE	History Sync for Fitbit to Apple Health #fitbit #getfit #workout #fitfam https://t.co/ki5D3SvDAd https://t.co/6hOCktxQLH
18	TRUE	#GetFit #Fitbit The Obesity-Infertility Connection https://t.co/QIcfHnq1Vt #Dolphins #Workout
19	FALSE	RT @MiddletonWatch: #Fitbit #GetFit Trying to Lose Weight? Don't Fall for This https://t.co/cfUzYAPGUm #KateMiddleton #Diet
20	FALSE	#Fitbit #Watches #FitbitSurge Small Size black Case, BOX and MANUAL ONLY https://t.co/ghDbGjASBj #Super #Watch https://t.co/UFfKYgjcot
21	FALSE	<U+231A><ed><U+00A0><U+00BD><ed><U+00B1><U+008D>#Fitbit #Watches #FitbitSurge #Fitness Activity Tracker Superwatch with Heart Rate Monitor (Small)... https://t.co/EFyu9rlehl
22	FALSE	#Health #Solution Ease Into Your Diet For Stress Free FALSE Success https://t.co/gAVswcyPT1 #Cavaliers #FitBit
23	FALSE	<U+231A><U+231A><ed><U+00A0><U+00BD><ed><U+00B1><U+008D>#Fitbit #Watches #FitbitSurge Large Black Used #Super #Watch https://t.co/S2fRnxfENS https://t.co/Ha2yS24rQ7... https://t.co/DXScYL3yIS
24	TRUE	I send one little "Neener-Neener" text to my dad about having more #fitbit steps and now he's doing 20K a day??
25	TRUE	#Fitbit #GetFit Defeating Emotional Eating https://t.co/NiDF39LlaN #KateMiddleton #Diet
26	FALSE	History Sync for Fitbit to Apple Health #fitbit #gym #apple #fitness #iphone #health https://t.co/ki5D3SvDAd https://t.co/ICRzkIbrm1

This random sample served as the training set in developing a predictive model through artificial intelligence. We coded the random sample in an Excel spreadsheet and used the command **read.csv()** in R. We maintained the corpus created and transformations achieved previously during the data preprocessing phase. We then created two more document term matrices using the random sample and removed insignificant terms. Data preparation was required to transform the document term matrix into a table, and then into a data frame, which allows for the building of a model, known as artificial intelligence.

The first step in building a model involved installing and loading the package "*e1071*" into the current R session. The model created was done with the support vector regression function, **svm()**. Once fitting the model, R generated its summary, which determined the total accuracy of the model, and its confusion matrix, which counts the number of times the predicted category mapped to the various true categories. The installation and running of package "*ROCR*" generated a visual of the predicted model created.

After the model was created, we deployed the model to a larger set of Fitbit tweets; the new Fitbit file contained five thousand tweets, rather than all twenty thousand, due to the capabilities of the machine running R. The new Fitbit file was read by R, a corpus created, the data transformed, a document term matrix created and transformed into a data frame to prepare the data for model deployment. Once the data was prepared, the model was deployed to the five thousand Fitbit tweets. R generated an Excel file with the actual predictions of the five thousand tweets as seen in Figure 3. The Excel spreadsheet contains one column of true or false values that represent the tweet's ability to relate to fitness.

Figure 3: Fitbit Model Actual Predictions

	A	B
1	Actualprediction	
2	FALSE	
3	TRUE	
4	FALSE	
5	FALSE	
6	TRUE	
7	TRUE	
8	TRUE	
9	FALSE	
10	FALSE	
11	FALSE	
12	FALSE	
13	FALSE	
14	FALSE	
15	FALSE	
16	FALSE	
17	FALSE	
18	FALSE	
19	FALSE	
20	FALSE	
21	FALSE	
22	FALSE	
23	FALSE	
24	FALSE	
25	FALSE	
26	FALSE	

The supervised learning technique of creating artificial intelligence was only applied to the Fitbit data due to the quantity of tweets gathered during the study. The R code used throughout the process of artificial intelligence is included in Appendix A, Section 4.

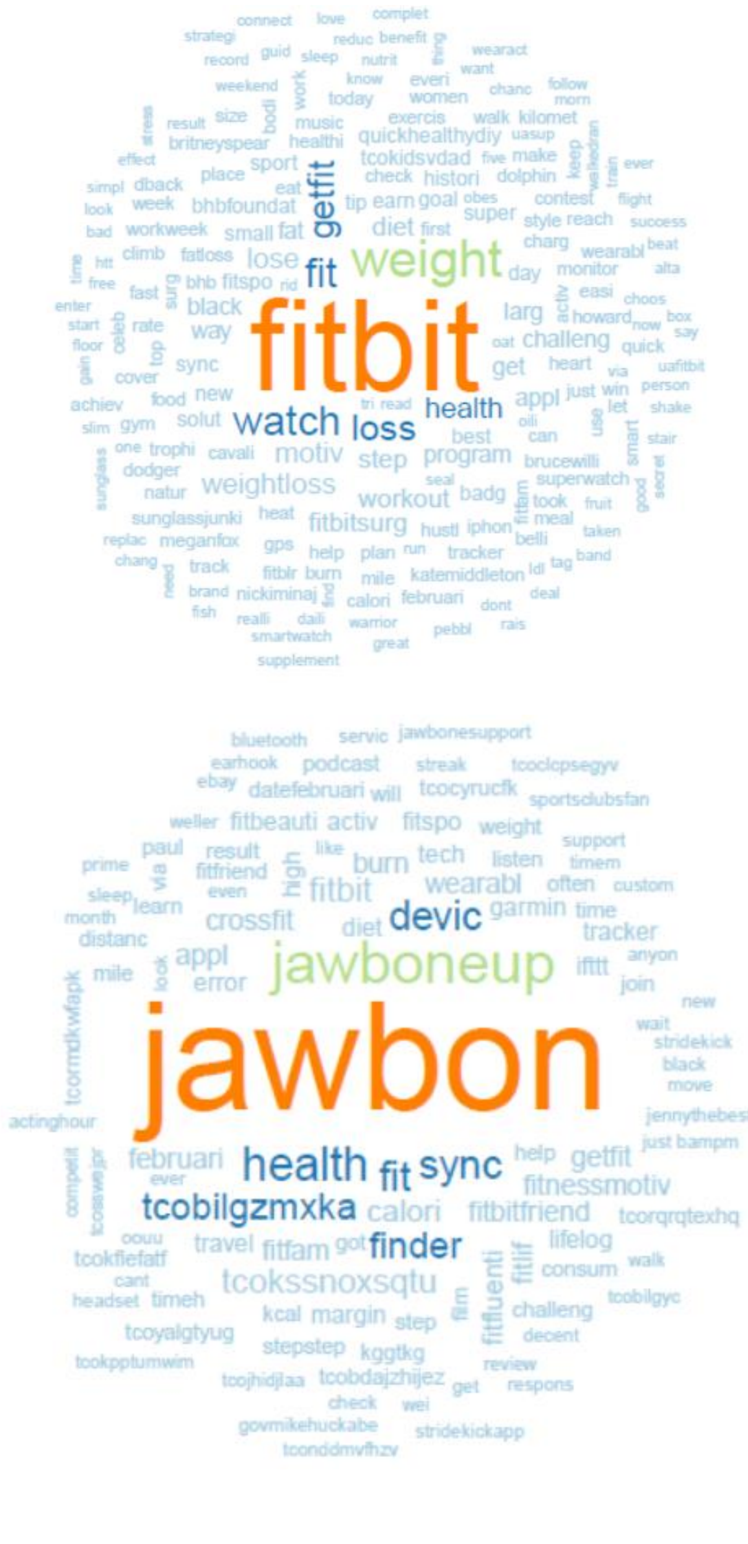
Results

The following sections display the results gathered from the research methodology stated above. Displayed first are the word clouds generated by R during the topic modeling process. Each of the subsequent word clouds, demonstrated in Figure 4, were developed based on the corpus of each data set. The word clouds expose the frequent words present among the collection of tweets posted about each company. The corpus visuals are displayed in the following order: Fitbit, Jawbone, Misfit, Withings.

The second portion of the results, presented in Figure 5, list the topics generated when topic modeling was applied to the data sets using R. For each company, four different tables are provided, indicating the variety in results for each of the two algorithms implemented, VEM and Gibbs, and for the impact of the tf-idf transformations on the data. Displayed alongside the tables are the associated word clouds for each of the tables presented, depicting the information in a visual form. Figure 5 shows the results in the following order: Fitbit, Jawbone, Misfit, Withings.

The final component of the research results is demonstrated in Figure 6. Included in the results are the R output values related to the summary statistics of the model created in developing artificial intelligence for the Fitbit tweets and the model's confusion matrix.

Figure 4: Corpus Word Clouds




misfit
exercis
fit
workout
getfit
flash
flashi
get
monitor
link
color
love
sleep
ever
best
check
track
ray
yet
yes
thompsonsgarag
digitalnomad
wunderkind
touchscreen
tweet
weekend
teamthompson
soundcloud
threaten
threichmil
montelshous
thompson
tonight
launch
iamfonda
smartwatch
muaythai
podcast
fighter
yesterday
eduauboedubeub
downeduaubdedubueduaubdedubuaeduaubdedubuafgawvi

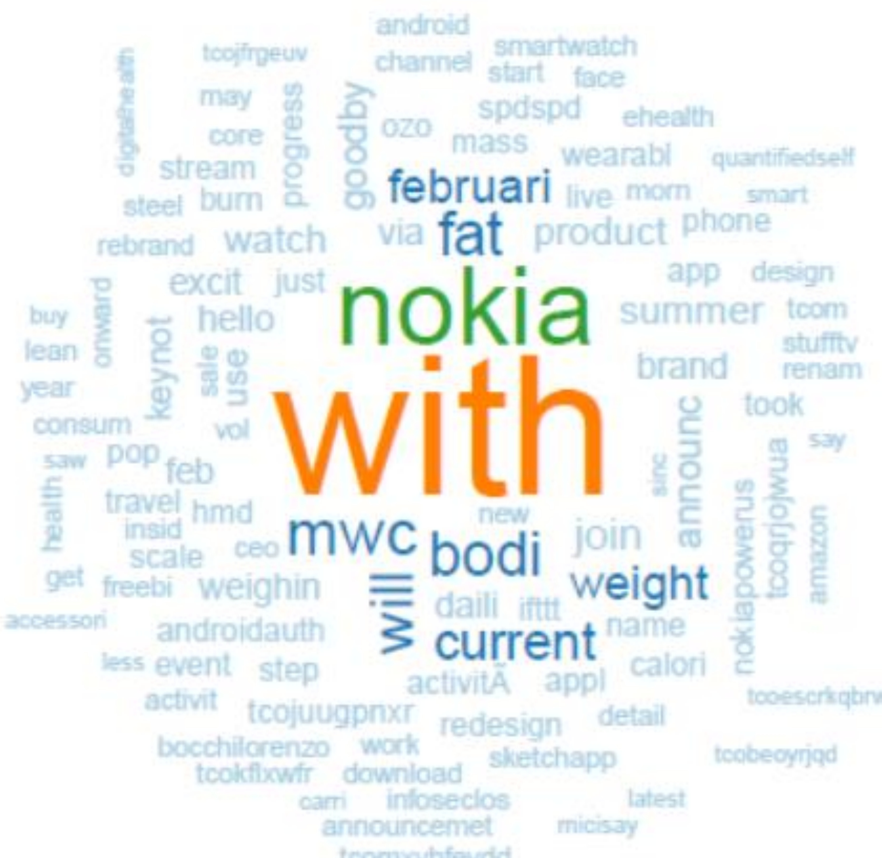

with
nokia
fat
februari
mwc
bodi
weight
current
will
goodby
keynot
announc
android
smartwatch
channel
start
face
spdspd
ehealth
ozo
mass
wearabl
quantifiedself
stream
progress
steel
burn
live
morn
smart
rebrand
watch
via
product
phone
excit
just
app
design
hello
summer
tcom
stufftv
brand
renam
took
nokiapowerus
amazon
weighin
androidauth
event
step
activit
redesign
sketchapp
bocchilorenzo
download
infoseclos
announcemet
latest
micisay

Figure 5: Topic Modeling Lists and Associated Word Clouds

Fitbit VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	fitbit	fitbit	weight	fitbit	fitbit
2	earn	health	weightloss	getfit	get
3	badg	appl	fitbit	weight	challeng
4	health	sync	loss	lose	step
5	solut	histori	fit	motiv	goal

Fitbit VEM Model (td-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	weight	weight	fit	getfit	watch
2	fit	watch	weight	fit	motiv
3	loss	loss	workout	weight	weight
4	lose	step	lose	loss	super
5	health	weightloss	getfit	health	program
6	diet	goal	diet	get	health

Fitbit Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	fitbit	watch	fitbit	weight	fitbit
2	health	fit	weightloss	loss	step
3	workout	fitbitsurg	program	motiv	get
4	getfit	larg	fat	lose	challeng
5	appl	black	bhbfoundat	diet	earn

Fitbit Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	weight	getfit	fit	step	get
2	loss	health	watch	earn	fat
3	motiv	lose	fitbitsurg	badg	challenge
4	weightloss	workout	larg	charg	goal
5	diet	appl	black	sunglassjunki	day
6	program	way	super	burn	bhbfoundat

Jawbone VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	jawbon	jawbon	jawbon	jawbon	health
2	health	film	ifttt	got	jawbon
3	sync	paul	diet	get	fitbit

Jawbone VEM Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	calori	devic	fitbit	health	consum
2	februari	jawboneup	health	jawboneup	via
3	wearabl	finder	fitbit	sync	film

Jawbone Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	jawbon	jawbon	fit	jawbon	health
2	devic	calori	fitbit	jawboneup	sync
3	finder	februari	wearabl	fitlif	jawboneup

Jawbone Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	health	devic	calori	fit	sync
2	fitbitfriend	finder	burn	fitbit	jawboneup
3	jawboneup	tcokssnoxsqtu	februari	wearabl	tcobilgzmxka

Misfit VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	misfit	misfit	fit	misfit	misfit
2	thompsonsgarag	golang	misfit	link	goal
2	come	like	get	exercis	activ
4	montelshous	thompsonsgarag	sleep	fit	like
5	weekend	digitalnomad	yet	ever	face

Misfit VEM Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	first	fit	fit	link	tonight
2	golang	get	get	exercise	thompsonsgarag
3	new	flash	sleep	check	montelshous
4	activ	love	yet	fit	daili
5	goal	color	yes	ever	goal

Misfit Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	exercis	misfit	get	misfit	misfit
2	workout	like	fit	fit	tonight
3	link	thompsonsgarag	getfit	sleep	can
4	ever	new	monitor	yet	daili
5	track	golang	flash	yes	via

Misfit Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	thompsonsgarag	exercis	fit	fit	get
2	tonight	link	exercis	workout	fit
3	like	ever	sleep	goal	getfit
4	detlefmus	check	ray	monitor	flash
5	time	best	yes	daili	love

Withings VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	with	with	with	with	with
2	live	activitA	mwc	bodi	nokia
3	nokia	pop	nokia	Fat	summer
4	will	step	will	current	product
5	use	via	announc	weight	brand

Withings VEM Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	fat	nokia	nokia	nokia	nokia
2	bodi	live	will	mwc	summer
3	current	will	mwc	goodby	product
4	februari	redesign	announc	hello	will
5	weight	use	join	brand	wearable

Withings Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	bodi	use	nokia	with	with
2	fat	feb	mwc	product	nokia
3	current	app	will	via	summer
4	weight	just	join	activitA	brand
5	februari	hmd	announc	scale	goodby

Withings Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	mwc	nokia	bodi	februari	use
2	will	product	fat	weighin	feb
3	join	summer	current	via	hmd
4	announc	brand	weight	activitA	scale
5	excit	mwc	watch	will	nokiapowerus

Figure 6: Artificial Intelligence Model

Figure 6A: Model Summary

```
Call:
svm(formula = Fitness ~ ., data = dfc, type = "C-classification", kernel = "radial", cost = 7, gamma = 0.05, cross = 10, cachesize = 1000, probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  7
      gamma:  0.05

Number of Support Vectors:  161

 ( 42 119 )


Number of Classes:  2

Levels:
 FALSE TRUE

10-fold cross-validation on training data:

Total Accuracy: 84
Single Accuracies:
 80 85 85 65 90 80 90 85 95 85
```

Figure 6B: Confusion Matrix

```
> pred <- fitted(model); y <- dfc$Fitness; table(pred, y)
       y
pred    FALSE TRUE
  FALSE   142    0
  TRUE      2   56
```

Discussion

I. Topic Modeling

The topic modeling results generated by R revealed the underlying topics that exist within the datasets analyzed. The word cloud visuals provide high-level overviews of the terms contained in the document term matrices and allowed researchers to see the significant keywords derived from the algorithms and transformations performed.

The topic modeling tables contributed to a more comprehensive understanding of the data. For each list of topic terms within the tables, researchers attempted to summarize the list of terms with a common theme. Topic themes included summaries such as weight loss, daily fitness, health and motivation, wearable technology, and burning calories. The entire set of tables with summary rows are provided in Appendix B.

The summary analysis results offered limited variation from one company to the next. Of the four companies, the Fitbit tweets provided the most value to the analysis due to the quantity of tweets that were posted and gathered during the collection period. Although the tweets regarding Jawbone, Misfit, and Withings reflected similar trends as those of Fitbit, many of the topics presented challenges in creating a summary due to the presence of insignificant terms. Such terms could often be found as hashtag terms in the original tweets. Of the 80 total topic summaries, 17 of the topics, roughly 21 percent, were unable to be classified. The 17 unclassified topics came from the companies other than Fitbit. All twenty Fitbit topics were easily characterized into a related theme.

Based on the analysis completed, the most significant research development involved the prevalence of weight loss and related topics contained in the tweets posted by users of Fitbit, Jawbone, Misfit, and Withings wearable technology devices. Of the 20 Fitbit topic summaries, 7 topic summaries included the term weight loss and no other term surfaced as many as or more than 7 times within the Fitbit data. Although in a less significant manner, terms related to weight loss surfaced in the topic modeling of Jawbone, Misfit, and Withings user tweets.

II. Artificial Intelligence

The creation of artificial intelligence was applied to only the Fitbit dataset due to the insufficient number of tweets gathered from the three competing companies. A model was created with the capability of categorizing tweets into different groups based on the patterns that the machine learned during the training process. As tweets are posted, they can be classified automatically with the use of artificial intelligence.

Fitbit's artificial intelligence model had a total accuracy of 84 percent, which was deemed acceptable by researchers striving for an accuracy greater than 80 percent. To further analyze the capability of the model, researchers ran a confusion matrix, which describes the performance of a classification model based on a set of training data. The matrix determined that of the 200 training tweets, the classifier predicted “false,” as it relates to a tweet’s relation to fitness, 142 times and “true” 56 times. The model only incorrectly classified two tweets; it classified two tweets that had been previously coded as not relating to fitness as tweets that related to fitness. With an error percentage of only one percent, researchers were satisfied with the performance of the model.

III. Research Limitations

Limitations existed within the research performed for the purpose of the thesis. One limitation involved R’s search application program interface, API, which allowed the collection of data within a period of ten days as of the date the data was collected. The capacity constraint impacted the number of tweets capable of being gathered and analyzed for Jawbone, Misfit, and Withings. It also affected the quality of tweets gathered; had researchers been able to gather samples of tweets throughout the year, they would have gathered a more representative sample of the content that users typically post tweets about. With such an insignificant amount of tweets collected from these three companies, researchers relied on the data collected from Fitbit for their analysis and conclusions.

Another limitation of the study involved the use of the tweet search criteria. The research required the analysis of tweets posted by product users of each company. To accomplish this, researchers collected data using the criteria of a hashtag followed by the company name. Researchers assumed these tweets were posted by users of the company’s products; however, any Twitter user has the capability to tweet #Fitbit, #Jawbone, #Misfit, #Withings even if they are not users of the company’s products.

Conclusion

Initial research of the application of wearable technology in the insurance industry revealed that significant opportunity exists for the benefit of both the policyholders and the insurance companies. Although some companies have introduced programs for their employees or for a small population of policyholders, complete integration has yet to become a widely accepted standard. The research performed aimed to identify the policyholders that insurance companies should target premium discounts toward should it be determined that discounts are necessary.

The research conducted suggests that insurance companies may want to consider lowering premiums for policyholders who track their health data with wearable technology. More specifically, insurance companies could reduce premiums for those who use wearable technology to increase their level of physical activity with the attempt to lose weight. Lack of sufficient physical activity puts individuals at risk for a variety of chronic diseases including cardiovascular disease, cancer, and obesity. If policyholders increased the amount of physical activity they regularly participate in, their risk of premature death would be reduced by 20 to 35 percent (Warburton, Nicol, & Bredin, 2006).

In addition, obesity is among the leading causes of preventable death in the United States. The condition affects more than one-third of adults and it costs hundreds of billions of dollars each year (Adult Obesity Facts, 2016). To combat the obesity epidemic, insurance companies could reduce premiums for those using wearable technology to increase their level of physical activity and lose weight. In doing this, the company's life insurance policyholders will be healthier, thus decreasing the chance of the policyholder filing a claim. The reduced number of claims will result in significant cost savings for the insurance companies.

References

Adult Obesity Facts. (2016, September 01). Retrieved May 05, 2017, from https://www.cdc.gov/obesity/data/adult.html

Bort, J. (2014, July 07). This Company Saved A Lot Of Money By Tracking Their Employees With Fitbits. Retrieved April 25, 2017, from http://www.businessinsider.com/company-saved-money-with-fitbits-2014-7

Crawford, K., Lingel, J., & Karppi, T. (2015). Our metrics, ourselves: A hundred years of self-tracking from the weight scale to the wrist wearable device. European Journal of Cultural Studies, 18(4-5), 479-496. doi:10.1177/1367549415584857

de Waal, M. J. (2016). The Good, Bad, and Ugly For Insurance and Wearable Tech. Insurance Networking News, 19(1), 17.

Fitbit. (2017). Retrieved May 02, 2017, from https://www.fitbit.com/about

Jawbone® | About The Company. (2017). Retrieved May 02, 2017, from https://jawbone.com/about

John Hancock's Plan: Use Fitbit Wearables to Adjust Life Premiums. (2015). Insurance Networking News, 18(2), 8.

Kaewkannate, K., & Kim, S. (2016). A comparison of wearable fitness devices. BMC Public Health, 16(1), 1-16. doi:10.1186/s12889-016-3059-0

Misfit: Fitness Trackers & Wearable Technology. (2017). Retrieved May 02, 2017, from https://misfit.com/

Pitsker, K. (2016). Watching Every Step You Take. Kiplinger's Personal Finance, 70(10), 70.

Soderland, S. (2015, September 14). What the Insurance Industry Can Do to Fix Health Care. Retrieved May 05, 2017, from https://hbr.org/2014/12/what-the-health-insurance-industry-can-do-to-fix-health-care

Warburton, D. E. R., Nicol, C. W., & Bredin, S. S. D. (2006). Health benefits of physical activity: the evidence. CMAJ : Canadian Medical Association Journal, 174(6), 801–809. http://doi.org/10.1503/cmaj.051351

What is R? (n.d.). Retrieved May 01, 2017, from https://www.r-project.org/about.html

Withings. (2017). Retrieved May 02, 2017, from https://www.withings.com/fr/en/about-withings

Appendix A: R Code

A.1 Data Collection

Data Collection.R*

```r
#First need to create an App at https://dev.twitter.com/

#Installing necessary packages
#install.packages("devtools")
#install.packages("httr")
#install.packages("twitteR")

library(devtools)
library(httr)
#install_github("twitteR", username = "geoffjentry")
#install.packages('base64enc')

library(twitteR)
#library(base64enc)
setup_twitter_oauth("ujtzSV9DpGGHBgLRIphMN5UDO", "sVLHp5p8fDsvfdbjehPn1d2KpdUKDwMlcC4QkZv5OMTdZ3IC6v",
                    "572270706-tpneOe6wwFjeqXUPzRMVbIWsolAIjBcGIgCIMV3i", "8cjNeaguVr7uBzDHI96OFWiwr5uOYnOxmWAqLFjjBBH3N")

#------Collecting Tweets-----------
tweets <- searchTwitter("#fitbit", n=90000, lang="en", since = "2016-01-01")
tweets.df <- twListToDF(tweets)
write.csv (tweets.df, file='C:/Users/Kaleigh/Documents/UNH/Spring 2017/Thesis/Data Collection/Fitbit.csv',row.names=F)

```

22:1 Collecting Tweets ÷ R Script ÷

A.2 Data Preprocessing

Fitbit Data PreProcessing.R*

```r
#install.packages("tm")
#install.packages("Snowballc")
#install.packages("slam")
library(tm)
library(SnowballC)
FitbitTweets <- read.csv(file='C:/Users/Kaleigh/Documents/UNH/Spring 2017/Thesis/texts/Fitbit.csv', head=TRUE)

#Creating corpus
corpus <- Corpus(VectorSource(FitbitTweets$text))

#Begin transformations
removeUnprintable <- function(x) gsub("[^[:alnum:]///' ]", "", x)
corpus <- tm_map(corpus, content_transformer(removeUnprintable))
removeURLs <- function(x) gsub("http[[:alnum:]]*", "", x)
corpus <- tm_map(corpus, content_transformer(removeURLs))
removeAccts <- function(x) gsub("@[[:alnum:]]*", "", x)# added this line to remove account names
corpus <- tm_map(corpus, content_transformer(removeAccts))# and this line as well.
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, stripWhitespace)
myStopWords <- c(stopwords('english'), "amp", "rt", "gt")
corpus <- tm_map(corpus, removeWords, myStopWords)
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, stemDocument)
corpus <- tm_map(corpus, stripWhitespace)

#Ensure transformations were done correctly
corpus[[1]]$content
corpus[[2]]$content
corpus[[3]]$content
corpus[[4]]$content
corpus[[5]]$content

```

34:1 (Top Level) ÷ R Script ÷

A.3 Unsupervised Learning: Topic Modeling

A.3.1 Data Preparation and Topic Modeling

```r
Fitbit Topic Modeling.R*

library(slam)
dtm <- DocumentTermMatrix(corpus)
summary(col_sums(dtm)) #Summary of term-frequency (tf) scores
term_tfidf <- tapply(dtm$v/row_sums(dtm)[dtm$i], dtm$j, mean) * log2(nDocs(dtm)/col_sums(dtm > 0)) #tf-idf transformation
summary(term_tfidf) #Summary of tf-idf scores

dtm1 <- dtm[,term_tfidf >= .10] #Removing terms whose tf-idf score doesn't meet the criterion.
dim(dtm) #Checking the number of rows and columns in dtm
dim(dtm1) #Checking the number of rows and columns in dtm 1
dtm2 <- dtm1[row_sums(dtm1) > 0,] #Removing rows that contain only zeros as a result of the previous line
dim(dtm1) #Checking the number of rows and columns in dtm 1
dim(dtm2) #Checking the number of rows and columns in dtm 2
summary(col_sums(dtm2))

library("topicmodels")
#-------------VEM--------------
Fitbit_VEM <- LDA(dtm, 8, method = "VEM", control = NULL, model = NULL) #Building a topic model with the original term-doc matrix. Generates 5 topics.
Terms <- terms(Fitbit_VEM, 6) #Shows 6 terms in each topic.
Terms[, 1:5] #Shows 5 topics generated.

Fitbit_VEM2 <- LDA(dtm2, 5, method = "VEM", control = NULL, model = NULL) #Building a topic model with the tf-idf transformed term-doc matrix with highly-frequent terms removed.
Terms2 <- terms(Fitbit_VEM2, 6)
Terms2[, 1:5]

#--------Gibbs-----------
Fitbit_Gibbs <- LDA(dtm, 5, method = "Gibbs", control = NULL, model = NULL)
Terms3 <- terms(Fitbit_Gibbs, 5)
Terms3[, 1:5]

Fitbit_Gibbs2 <- LDA(dtm2, 5, method = "Gibbs", control = NULL, model = NULL)
Terms4 <- terms(Fitbit_Gibbs2, 6)
Terms4[, 1:5]

#-------Assign a topic number to each tweet in the original tweet document--------
ldaOut.topics <- as.matrix(topics(Fitbit_Gibbs2))
write.csv(ldaOut.topics,file="Fitbit_Gibbs2_DocsToTopics.csv")

71:1  Assign a topic number to each tweet in the original tweet document   R Script
```

A.3.2 Word Cloud Generation

```r
#------Frequency------
#freq <- findFreqTerms(dtm, lowfreq = 20)
#write.csv(freq[1:6108], "Fitbit_freq.csv")


#Word Cloud Generator
#install.packages("wordcloud")
#install.packages("RColorBrewer")
library("wordcloud")
library("RColorBrewer")

wordcloud(dtm, max.words = 200, random.order = FALSE, colors=brewer.pal(8, "Paired"))

71:1  Assign a topic number to each tweet in the original tweet document   R Script
```

A.4 Supervised Learning: Artificial Intelligence

A.4.1 Random Sample and Data Preparation

```r
Supervised Learning.R*

#Moved the @ account name identifier removal before punctuation to see if we can remove more account names.
library(tm); library(SnowballC)
FitbitRandomSample <- read.csv(file='/Users/Kaleigh/Documents/UNH/Spring 2017/Thesis/Supervised Learning/Fitbit_Random_Sample.csv', head=TRUE)

#Creating corpus
corpus <- Corpus(VectorSource(FitbitRandomSample$Text))

#Begin transformations
removeUnprintable <- function(x) gsub("[^[:alnum:]///' ]", "", x)
corpus <- tm_map(corpus, content_transformer(removeUnprintable))
removeURLs <- function(x) gsub("http[[:alnum:]]*", "", x)
corpus <- tm_map(corpus, content_transformer(removeURLs))
removeAccts <- function(x) gsub("@[[:alnum:]]*", "", x)#Added this line to remove account names
corpus <- tm_map(corpus, content_transformer(removeAccts))#And this line as well.
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, stripWhitespace)
myStopWords <- c(stopwords('english'), "amp", "rt", "gt")
corpus <- tm_map(corpus, removeWords, myStopWords)
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, stemDocument)
corpus <- tm_map(corpus, stripWhitespace)

#To dtm (document-term matrix)
doc_matrix <- DocumentTermMatrix(corpus, control = list(weighting =function(x) weightTfIdf(x, normalize = FALSE)))#Changing weightTfIdf
doc_matrix2 <- removeSparseTerms(doc_matrix, sparse=.98)#Changed sparse from .95 as this cut out everything.

#Data prep
d <- as.table(doc_matrix2) #Document-Term Matrix to table
df <- as.data.frame.matrix(d) #Table to Data Frame
dfc <- cbind.data.frame(FitbitRandomSample$Fitness, df) #Combining the class and the features (aka variables)
names(dfc)[names(dfc)=="FitbitRandomSample$Fitness"] <- "Fitness" #Change variable name. Having $ causes a problem.

34:1  (Top Level)   R Script
```

A.4.2 Artificial Intelligence Model Creation

```r

#Building a model (Artificial Intelligence)
#install.packages('e1071')
library(e1071)
model <- svm(Fitness ~ ., data = dfc,
             type = "C-classification",
             kernel = "radial",
             cost = 7,
             gamma = 0.05,
             cross = 10,
             cachesize = 1000,
             probability = TRUE)
summary(model)

#Confusion Matrix
pred <- fitted(model); y <- dfc$Fitness; table(pred, y)

#Performance: ROC and AUC
#producing ROC Curves
svm.prob <- predict(model, type="prob", newdata=df, probability = TRUE)
#install.packages('ROCR')
library(ROCR)
pred1 <- prediction(attr(svm.prob, "probabilities")[,2], y); perf <- performance(pred1, "tpr", "fpr"); plot(perf, colorize=T, lwd=3, main= "ROCR")
#AUC
auc <- performance(pred1, "auc"); auc
```

A.4.3 Artificial Intelligence Model Deployment

```r
#For this, we need model from Text Mining with SVM
#Actual Prediction

ds <- read.csv(file='/Users/Kaleigh/Documents/UNH/Spring 2017/Thesis/Supervised Learning/Fitbit New.csv', head=TRUE)

#Creating corpusA
corpusA <- Corpus(VectorSource(ds$text))

# begin transformations
removeUnprintable <- function(x) gsub("[^[:alnum:]///' ]", "", x)#this was the culprit!!!
corpusA <- tm_map(corpusA, content_transformer(removeUnprintable))
removeURLs <- function(x) gsub("http[[:alnum:]]*", "", x)
corpusA <- tm_map(corpusA, content_transformer(removeURLs))
removeAccts <- function(x) gsub("@[[:alnum:]]*", "", x)# added this line to remove account names
corpusA <- tm_map(corpusA, content_transformer(removeAccts))# and this line as well.
corpusA <- tm_map(corpusA, removeNumbers)
corpusA <- tm_map(corpusA, removePunctuation)
corpusA <- tm_map(corpusA, content_transformer(tolower))
corpusA <- tm_map(corpusA, stripWhitespace)
myStopwords <- c(stopwords('english'), "amp", "rt", "gt")
corpusA <- tm_map(corpusA, removeWords, myStopwords)
corpusA <- tm_map(corpusA, stripWhitespace)
corpusA <- tm_map(corpusA, stemDocument)
corpusA <- tm_map(corpusA, stripWhitespace)

#To dtm (document-term matrix)
doc_matrixA <- DocumentTermMatrix(corpusA, control = list(weighting =function(x) weightTfIdf(x, normalize = FALSE)))

#To data frame
d_Pred <- as.table(doc_matrixA) #Document-Term Matrix to table
dfPrediction <- as.data.frame.matrix(d_Pred) #Table to Data Frame

#Data prep
fnames <- colnames(df) #pull feature names from the training dataset
common_columns <- intersect(fnames, colnames(dfPrediction)) #findings columns (i.e., features) of the actual dataset that also exist in the predictive model.
subset_Pre <- dfPrediction[common_columns] #creating a sub-dataset with the common features
different_cols <- setdiff(fnames, common_columns) #finding features that exist in the predictive model but not in the actual dataset.
subset_Pre[,different_cols] <- 0 #adding those features to the sub-dataset. The # of columns is always the same as fnames.

#Acutal Prediction
Actualprediction <- predict(model, subset_Pre) #here is where we use the predictive model.
p = data.frame (Actualprediction)
write.csv (p, file='/Users/Kaleigh/Documents/UNH/Spring 2017/Thesis/Supervised Learning/ActualPredictions.csv', row.names=F)
```

Appendix B: Topic Modeling Analysis

B.1 Fitbit

Fitbit VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	fitbit	fitbit	weight	fitbit	fitbit
2	earn	health	weightloss	getfit	get
3	badg	appl	fitbit	weight	challeng
4	health	sync	loss	lose	step
5	solut	histori	fit	motiv	goal
Summary	**Product Features**	**Health Tracking**	**Weight Loss**	**Weight Loss Motivation**	**Daily Fitness**

Fitbit VEM Model (td-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	weight	weight	fit	getfit	watch
2	fit	watch	weight	fit	motiv
3	loss	loss	workout	weight	weight
4	lose	step	lose	loss	super
5	health	weightloss	getfit	health	program
6	diet	goal	diet	get	health
Summary	**Weight Loss / Diet**	**Weight Loss / Activity**	**Fitness**	**Fitness / Health**	**Health / Motivation**

Fitbit Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	fitbit	watch	fitbit	weight	fitbit
2	health	fit	weightloss	loss	step
3	workout	fitbitsurg	program	motiv	get
4	getfit	larg	fat	lose	challeng
5	appl	black	bhbfoundat	diet	earn
Summary	**Fitness**	**Products**	**Weight Loss Program**	**Diet / Weight Loss**	**Product Features**

Fitbit Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	weight	getfit	fit	step	get
2	loss	health	watch	earn	fat
3	motiv	lose	fitbitsurg	badg	challenge
4	weightloss	workout	larg	charg	goal
5	diet	appl	black	sunglassjunki	day
6	program	way	super	burn	bhbfoundat
Summary	**Weight Loss Program**	**Health / Exercise**	**Product / Features**	**Product Features**	**Daily Fitness**

B.2 Jawbone

Jawbone VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	jawbon	jawbon	jawbon	jawbon	health
2	health	film	ifttt	got	jawbon
3	sync	paul	diet	get	fitbit
Summary	**Health Tracking**	–	**Fitness Time**	**Achievements**	**Health Tracking**

Jawbone VEM Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	calori	devic	fitbit	health	consum
2	februari	jawboneup	health	jawboneup	via
3	wearabl	finder	fitbit	sync	film
Summary	**Burn Calories**	**Product Features**	**Health Tracking**	**Health Tracking**	–

Jawbone Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	jawbon	jawbon	fit	jawbon	health
2	devic	calori	fitbit	jawboneup	sync
3	finder	februari	wearabl	fitlif	jawboneup
Summary	**Wearable Technology**	**Product Features**	**Wearable Technology**	**Fitness**	**Health Tracking**

Jawbone Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	health	devic	calori	fit	sync
2	fitbitfriend	finder	burn	fitbit	jawboneup
3	jawboneup	tcokssnoxsqtu	februari	wearabl	tcobilgzmxka
Summary	**Product Features**	–	**Burn Calories**	**Wearable Technology**	–

B.3 Misfit

Misfit VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	misfit	misfit	fit	misfit	misfit
2	thompsonsgarag	golang	misfit	link	goal
2	come	like	get	exercis	activ
4	montelshous	thompsonsgarag	sleep	fit	like
5	weekend	digitalnomad	yet	ever	face
Summary	–	–	**Fitness / Sleep Tracking**	**Exercise / Fitness**	**Activity Tracking**

Misfit VEM Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	first	fit	fit	link	tonight
2	golang	get	get	exercise	thompsonsgarag
3	new	flash	sleep	check	montelshous
4	activ	love	yet	fit	daili
5	goal	color	yes	ever	goal
Summary	**Activity Goal**	**Fitness / Products**	**Fitness / Sleep**	**Fitness / Exercise**	–

Misfit Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	exercis	misfit	get	misfit	misfit
2	workout	like	fit	fit	tonight
3	link	thompsonsgarag	getfit	sleep	can
4	ever	new	monitor	yet	daili
5	track	golang	flash	yes	via
Summary	**Fitness Tracking**	–	**Fitness**	**Fitness / Sleep**	–

Misfit Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	thompsonsgarag	exercis	fit	fit	get
2	tonight	link	exercis	workout	fit
3	like	ever	sleep	goal	getfit
4	detlefmus	check	ray	monitor	flash
5	time	best	yes	daili	love
Summary	–	**Exercise**	**Exercise / Sleep**	**Fitness Tracking**	**Fitness / Products**

B.4 Withings

Withings VEM Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	with	with	with	with	with
2	live	activitA	mwc	bodi	nokia
3	nokia	pop	nokia	fat	summer
4	will	step	will	current	product
5	use	via	announc	weight	brand
Summary	–	**Activity**	–	**Body Weight**	**Products**

Withings VEM Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	fat	nokia	nokia	nokia	nokia
2	bodi	live	will	mwc	summer
3	current	will	mwc	goodby	product
4	februari	redesign	announc	hello	will
5	weight	use	join	brand	wearable
Summary	**Body Weight**	**Products**	–	–	**Wearable Technology**

Withings Gibbs Model					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	bodi	use	nokia	with	with
2	fat	feb	mwc	product	nokia
3	current	app	will	via	summer
4	weight	just	join	activitA	brand
5	februari	hmd	announc	scale	goodby
Summary	**Body Weight**	**Application Use**	–	**Fitness Products**	–

Withings Gibbs Model (tf-idf)					
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	mwc	nokia	bodi	februari	use
2	will	product	fat	weighin	feb
3	join	summer	current	via	hmd
4	announc	brand	weight	activitA	scale
5	excit	mwc	watch	will	nokiapowerus
Summary	**Product**	**Health Tracking**	**Weight Management**	**Fitness Journey**	–