Mining Health-Related Issues in Consumer Product Reviews by Using Scalable Text Analytics

Manabu Torii, Sameer S. Tilak, Son Doan, Daniel S. Zisook and Jung-wei Fan

Medical Informatics, Kaiser Permanente Southern California, San Diego, CA, USA.

Supplementary Issue: Innovations in Clinical Informatics

ABSTRACT: In an era when most of our life activities are digitized and recorded, opportunities abound to gain insights about population health. Online product reviews present a unique data source that is currently underexplored. Health-related information, although scarce, can be systematically mined in online product reviews. Leveraging natural language processing and machine learning tools, we were able to mine 1.3 million grocery product reviews for health-related information. The objectives of the study were as follows: (1) conduct quantitative and qualitative analysis on the types of health issues found in consumer product reviews; (2) develop a machine learning classifier to detect reviews that contain health-related issues; and (3) gain insights about the task characteristics and challenges for text analytics to guide future research.

KEYWORDS: consumer health informatics, text mining, natural language processing, big data, online product reviews, syndromic surveillance

SUPPLEMENT: Innovations in Clinical Informatics

CITATION: Torii et al. Mining Health-Related Issues in Consumer Product Reviews by Using Scalable Text Analytics. *Biomedical Informatics Insights* 2016:8(S1) 1–11 doi: 10.4137/BII.S37791.

TYPE: Original Research

RECEIVED: February 15, 2016. **RESUBMITTED:** May 01, 2016. **ACCEPTED FOR PUBLICATION:** May 17, 2016.

ACADEMIC EDITOR: John P. Pestian, Editor in Chief

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,717 words, excluding any confidential comments to the academic editor.

FUNDING: The authors are salaried employees of Kaiser Permanente. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: jung-wei.x.fan@kp.org

Introduction

These days, much of our daily activities are digitized and recorded in some form. Analysis of such data presents the opportunity to gain insights into our daily activities, including health-related activities. Digitized records pertaining to health-related information can originate from various sources. The primary source of such information has been software systems that are designed to collect health-related information, such as electronic medical records. There has been vibrant data/text mining research on such formal health information.[1,2] Although not created for collecting health information, blogs and other social media have recently emerged as another unique complementary source of health-related information.[3–6] In this study, we focus on health-related information in consumer product reviews, which are still an underexplored data source of such information.

Many online vendors collect product ratings and reviews, which can serve as word-of-mouth advertising as well as feedback to the product manufacturer/merchant. As online consumer spending continues to grow,[7] the amount of consumer-generated reviews has also increased. Given the variety of goods sold online, it is not uncommon that consumer reviews discuss health-related issues. For example, a consumer may write a review about an adverse effect caused by the product or justification of choosing the product to avoid/alleviate a health issue, eg, "[This product is a] Major migraine trigger!" or "It's supposed to help literally pull gingivitis out." The number of reviews can easily reach millions on prominent retail sites, such as Amazon.com. Owing to advancements in software and hardware, processing of a huge dataset, which used to take weeks or months, can now be completed in hours or real time.

We were motivated to investigate health-related information in consumer product reviews based on two assumptions: (1) Given the vast amount of reviews available, we should be able to collect infrequent but valuable pieces of information by leveraging efficient big data techniques; (2) Even if the health issues mentioned in product reviews are not novel discoveries in themselves, it is still useful to summarize the different types and aspects of illness discussed on consumer products – ideally with discovering incidences/patterns to complement formally collected health data.

In this exploratory study, we conducted quantitative and qualitative analysis on the types of health issues found in consumer product reviews. We processed 1.3 million Amazon.com reviews on Grocery and Gourmet Food products using a scalable natural language processing (NLP) system based on Apache Unstructured Information Management Architecture (UIMA)[8] Asynchronous Scaleout (UIMA-AS). A subset of the concepts extracted were manually reviewed and annotated as relevant or irrelevant to health-related issues. With this dataset, a machine learning classifier was trained using

Apache Spark[9] for screening the relevant reviews. Descriptive statistics and manual inspection were conducted to analyze the results.

The three deliverables from this study were as follows: (1) quantitative and qualitative analysis on the types of health issues found in the reviews; (2) a machine learning classifier that can screen for reviews containing health-related issues; and (3) insights about the task characteristics and challenges for text analytics that will guide future research. In terms of practical impact, the study contributes to biomedical informatics by exploring the value of consumer product reviews as a complementary information source for the purpose of public health monitoring.

Background

An increasing number of studies are being published addressing mining of health-related information from nontraditional data sources. For example, Corley et al analyzed over 2 years' blog posts to detect influenza epidemic signals.[3] Ofoghi et al investigated the classification of emotion expressed in Twitter posts for disease outbreak detection and monitoring.[4] Aphinyanaphongs et al applied text classification for the detection of e-cigarette use and smoking cessation in Twitter.[5] Sarker et al.[6] conducted a literature survey on studies exploiting social media data for pharmacovigilance. Several publications have also focused on analyzing consumer reviews, such as automated summarization,[10] opinion/sentiment analysis,[11] and evaluation of the helpfulness of reviews.[12]

To our knowledge, however, there are few studies focusing on consumer product reviews as a source of mining health-related information. A recent study by Sullivan et al investigated adverse reaction to dietary supplements reported in Amazon user reviews of nutritional supplements.[13] The results suggest that product reviews can be an information source for monitoring adverse reactions reported on dietary supplements. Extrapolating these findings, mining a large set of reviews across many products can provide information on diverse health-related issues. Until recently, mining information from a large set of product reviews was difficult because of hardware and software limitations. Significant advancements in scalable analytics offers unprecedented computing efficiency at moderate cost[14–16] and enables large-scale data mining.

Material and Tools

Amazon reviews. Amazon.com is one of the major online retailers in the United States. In 2005, it had more than 304 million active customer accounts and $107 billion net sales.[17] Product reviews provided by customers available on the Amazon website contain highly valuable information. A recent CNET article writes: "Customer reviews have been a crucial part of Amazon's websites for over 20 years, with the written reviews and 5-star rating system becoming an important form of accountability and sign of popularity and quality for items buyers often cannot touch or test out before purchasing."[a]

In this study, we used customer product reviews on the Amazon.com website that had been previously obtained by McAuley et al.[18] The dataset was originally collected for data mining studies and then shared with the research community. The original collection contains 143.7 million product reviews between May 1996 and July 2014. The reviewed products, and hence the corresponding review text, are divided into 24 high-level categories, such as books, electronic, and movies and TV, as well as Grocery and Gourmet Food. The review and product information are stored in JSON[19] format, with fields containing review text, review date, product name, and product category, among other information. In our study, we used the category Grocery and Gourmet Food (or *grocery products* hereafter), which includes about 1.3 millions reviews on more than 120,000 products submitted by 774,000 users.

nQuiry and UIMA-AS. The nQuiry system is a comprehensive NLP pipeline developed by the Medical Informatics team of the Kaiser Permanente Southern California Medical Group. The nQuiry system uses Apache UIMA[8] and allows flexible decomposition of NLP tasks into modules. The core modules of the nQuiry system include tokenization, typo correction, sentence chunking, part-of-speech tagging, syntactic parsing, phrase extraction, concept candidate selection, concept searching, sense disambiguation, and negation/modality classification. The nQuiry system has been incorporated into several applications: automated diagnosis coding, evaluation and management coding, risk screening for aortic aneurysm, and cardiovascular risk factor identification.[20] The nQuiry deployment leverages the UIMA-AS framework. Multiple nQuiry processes are launched as UIMA-AS Service Instances on server machines, and they can be used in a parallel manner to handle a large collection of input text submitted by UIMA-AS Clients, where load balancing is facilitated by the asynchronous middleware using Apache ActiveMQ[21] implementation of Java Messaging Services. This framework can achieve high throughput by scaling out the workload linearly.

Apache Spark and MLlib. Big data technologies, such as Apache Hadoop[22] and Pig,[23] are powerful and convenient platforms for handling large datasets. We used Pig to build our post-NLP analytic pipeline. In Pig, the data flows are described using Pig Latin language, which then gets translated into MapReduce[24] jobs that exploit data parallelism. To facilitate analyzing millions of reviews processed by the nQuiry system, we used Hadoop SequenceFile – a flat file consisting of binary key/value pairs. In our case, we extracted filename/entities as the key/value pairs from a large number of outputs generated by the nQuiry system and aggregated them into several SequenceFiles, which could then be handled easily by calling user-defined functions in Pig Latin script.

[a]http://www.cnet.com/news/amazon-updates-customer-reviews-with-new-machine-learning-platform/

Apache Spark[25] is an open-source cluster computing framework that employs in-memory primitives for performance. Resilient Distributed Datasets are the key programming abstraction in Spark. Resilient Distributed Dataset is essentially a logical collection of data partitioned across machines that can be manipulated in parallel. Spark MLlib is a distributed machine learning framework on top of Spark Core. MLlib consists of common learning algorithms and utilities, including classification that we used in our study.

Methods

Data sampling and preprocessing. As an exploratory study, we chose the grocery products from the Amazon product categories that were available in the aforementioned dataset. This subset contains 1.3 million reviews that cover diverse product types, ranging from drinks and snacks to dietary supplements. We processed the reviews by the nQuiry system, followed by additional filters to narrow down selection of reviews containing health-related issues. Figure 1 shows an overview of the data processing workflow.

The following modules of the nQuiry system were essential in the first step of the workflow: tokenization, sentence chunking, part-of-speech tagging, syntactic parsing, phrase extraction, concept candidate selection, and concept searching. Specifically, the phrases extracted by the upstream modules were screened by the *concept candidate selector*, which preserved any phrase with headword determined as medically relevant. The screened candidate phrases were then looked up by the *concept searcher* into Unified Medical Language System (UMLS) concept/synonym indexes that had been created by using Apache Lucene.[26] The identified UMLS concepts were normalized with assignment to the corresponding Concept Unique Identifier (CUI). In our experiment, a total of 40 nQuiry engine instances were deployed as UIMA-AS Service Instances, and it took 6 hours to process the 1.3 million grocery product review texts. Over 2 million occurrences of UMLS concepts that belong to the DISO (Disorder) semantic group[27] were extracted from this dataset.

After applying the nQuiry system, as additional filtering of concept phrases detected, the *semantic type filter* was applied (Fig. 1). In this step, concept phrases associated with CUIs were filtered out unless the CUIs belong to the UMLS semantic types of Disease or Syndrome (T047) and Sign or Symptom (T184). The concept occurrences extracted from the grocery product dataset were reduced to 0.3 million through this step.

Corpus annotation. It was known that the concept phrases from the NLP and semantic type filtering still contained many false extractions. In order to further weed out the false extractions, we manually annotated a subset of the concept phrases for training a machine learning classifier that could determine whether a phrase really represented a health-related issue (Fig. 1). Since each review can contain more than one possible health issue, the annotation was performed on phrases, rather than on sentences or reviews. The granular unit of classification at the phrase level is more informative and can be easily interpreted at the review level if needed – that is, any review text that contains at least one *relevant* phrase will be considered as *health-related*. To be clear, the classification task referred hereafter is at the phrase level.

Three of the authors (MT, SD, and JF) read through ~1,700 phrases each (5,077 phrases in total) with given surrounding contexts and classified into two classes, *relevant* or *irrelevant*. Note that the annotations were done without considering CUI, because the manual decision implicitly involved disambiguation, ie, an incorrect sense (CUI) would result in classifying as irrelevant. As part of the process, tough cases were discussed among the annotators. In the end, it was agreed that our definition of relevance be made sensitive: any health issue that is related to humans and that concerns a consumer can be considered as relevant, so that both beneficial incidents (prevention/management) and detrimental ones (adverse effect) were included. The annotation also helped us understand the common types of errors by the NLP system.

After the entire dataset was annotated, 100 phrases annotated by each of the three annotators were annotated by the other two annotators (50 phrases each), and the total of

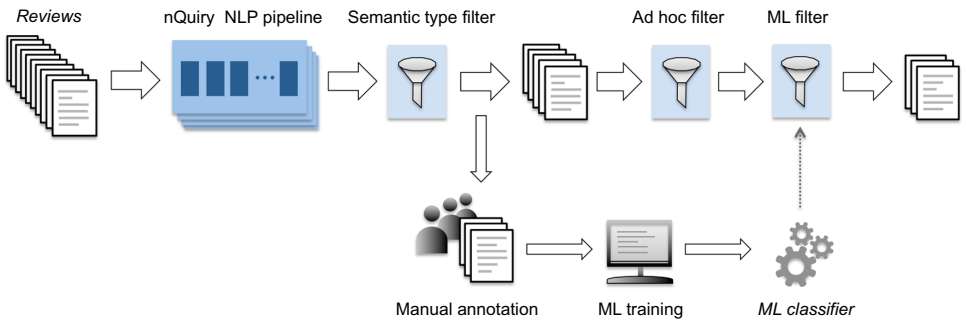

Figure 1. An overview of the data processing workflow.
Note: A clinical NLP pipeline, nQuiry, was used to extract phrases potentially relevant to health-related issues in consumer product reviews. Detected phrases were further filtered in the subsequent steps to narrow down the target concepts.

150 phrases were doubly annotated so that the agreement of the annotator pairs could be estimated. The agreement rate calculated for pairs of annotators was a mean Cohen's kappa of 0.751, which might be considered as *fair agreement* (between 0.67 and 0.8), but not necessarily *good* (above 0.8).[28] In terms of *F*-score, the agreement was calculated as 0.837 (see the next section on *F*-score calculation). Disagreement cases included difficult phrases, eg, "burns your mouth", which could simply mean spicy or spicy to the extent that it might be considered health-related. There were, however, several negated concepts among disagreed phrases, such as "it [coffee] doesn't give her acid reflux", which we had discussed not to annotate as relevant through our discussion but were found to be not consistently annotated in some cases.

As byproduct of the annotation, we identified phrases that were not of our interest but frequently detected by the nQuiry system, and incorporated ad hoc filters into the postprocessor of nQuiry. For example, after observing that the majority of occurrences of *hunger* and *thirst* in the product reviews did not provide valuable insight into consumer health, a decision was made to filter these phrases from nQuiry outputs.

Classifier training and testing. A machine learning classifier was developed to determine health relevance of candidate concept phrases identified by the nQuiry system. A Pig-based pipeline was created to process each phrase along with its context to generate the feature vector and train a classifier. Specifically, we used MLlib to train a classification model of logistic regression. A model was trained with standard feature scaling and L2 regularization. We initially tested classification algorithms other than logistic regression, such as support vector machine and multinomial and Bernoulli naïve Bayes, which are commonly used for text classification. Observing that our preliminary tests conformed to previously reported results on these classification methods (eg, logistic regression and support vector machine yielded comparable performance, and they outperformed naïve Bayes[29]), we decided to use logistic regression in our study, which can provide well-calibrated probability scores for predicted classes. When training a logistic regression model, we used the default parameters in MLlib and employed basic classification features, as discussed next, so that we could first establish a general workflow to analyze diverse reviews.

As for classification features, bags of words were used. This general approach was selected based on our experience in training similar classification models, specifically classifiers used in the nQuiry system for negation/modality detection.[30] Three bags of words were created from within-sentence context: (1) words left to the target phrase, (2) words right to the phrase, and (3) words in the phrase itself. A trained classifier predicts the relevance class with a probability score, and typically a threshold of 0.5 was used to differentiate the relevant class apart from the irrelevant class.

To study the sufficiency of training data size and also the appropriate ratio of relevant to irrelevant instances in the training dataset, we conducted two experiments with the classifier:

1. We varied the number of annotated instances in the training set so as to examine the impact of the training data size, ie, how the accuracy would improve as the training data size is increased. In particular, we were interested to see whether the improvement would hit plateau at a certain point, which indicates even adding more annotations would not significantly help the accuracy.
2. We varied (down-sampled) the number of irrelevant instances in the training data and examined how the ratio of the two classes would affect the classification accuracy. Note that in this experiment each classifier might have different optimal probability thresholds, given that they were trained with differently manipulated class distributions. Therefore, we computed the precision–recall curve as more objective evaluation to show the tradeoffs.

To obtain robust accuracy estimation in the experiments, we performed resampling evaluation tests, in which we randomly split the annotated training instances into training and testing subsets: 80% for training and 20% for testing. The process was repeated 50 times to compute averaged metrics of precision, recall, and *F*-score, as defined below.

Precision = True positive/(True positive + False positive)
Recall = True positive/(True positive + False negative)
F-score = 2 × Precision × Recall/(Precision + Recall)

After the above two experiments, we trained a final classifier using all the annotated data, which was found to have the best ratio of relevant to irrelevant instances as reported in the "Results" section. The final classifier was applied to screen for relevant concept phrases from a large unseen dataset of reviews processed by the nQuiry system. Similar to evaluating information retrieval systems, it was impractical to calculate the recall of the developed workflow on this large unseen dataset, besides that our interest in the study was to discover and examine any health-related information buried in customer review data. Therefore, on this large unseen dataset, we focused on the precision of extracting health-related information within the cases that were judged as relevant per classifier confidence. Instances assigned with prediction probability of 1.0 were collected, and 100 samples among them were manually reviewed. To further examine the classification results, the remaining predictions in the probability range of [0.0, 1.0] were divided into six bins: [0.0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], [0.8, 1.0], and [1.0, 1.0]. The precision of each probability stratum was estimated by manually reviewing 200 samples and annotating each as either true positive or false positive.

Content analysis of the reviews and predictions. To quantitatively summarize the health-related issues, we computed the most frequently mentioned diseases/symptoms in the reviews. For each major disease/symptom, a couple of top associated product types were also provided to demonstrate the

potential value of identifying such relations. To qualitatively analyze the review contents, we went over 100 random manual relevant annotations and categorized them based on the nature of the health issues. We also reviewed the false positive disease/symptom extractions by NLP to identify challenges and opportunities for future improvement.

Results

Classifier tuning and prediction accuracy. The experiments were conducted using the manually annotated dataset, 5,077 phrases/instances, of which 35% were annotated as relevant. Results of the three major evaluations are summarized as follows:

Effect of training data size. The first experiment (Fig. 2) showed that the *F*-score of classification steadily increased as the training data size increased. The slowdown of improvement with increase of training data suggests that capping man-hour investment should be considered as long as satisfactory accuracy has been achieved.

Effect of class ratio. The second experiment (Fig. 3) showed that slightly better performance could be achieved when using the training data with the ratio of irrelevant to relevant of 2:1 (the *Double* in Figure 3). Therefore, we used all the annotated instances that had the Double class ratio. With that ratio, the blue line in Figure 3 indicates that the precision is close to 0.8 while the recall reaches 0.8.

Precision of the fully trained classifier. The precision of the final classifier per probability stratum is shown in Figure 4. Overall, the classifier learned reasonably well to generate prediction probability that correlated with true reliability. For example, the highest probability bin had a precision of ~82% while the lowest probability bin only had ~37%.

Health-related issues found in the reviews. The most frequent health issues observed in the reviews are summarized in Table 1. While these health issues were frequently observed in the fully annotated dataset and also in the large dataset used for the final experiment, the frequency counts shown in the table were obtained in the former dataset as the health issues were manually confirmed in that dataset. The top issue *pain* is vague and indicates that consumer review language can be terse and unclear, or the location of pain needs to be inferred from the type of the product. In Table 2, we list some representative examples of product types associated with each of the top five health issues from Table 1. It can be noted that certain products (eg, ginger candy) were reported as beneficial in some reviews and detrimental in others. Based on 100 manually annotated relevant cases, the health issues were analyzed according to the nature of the issue and assigned into

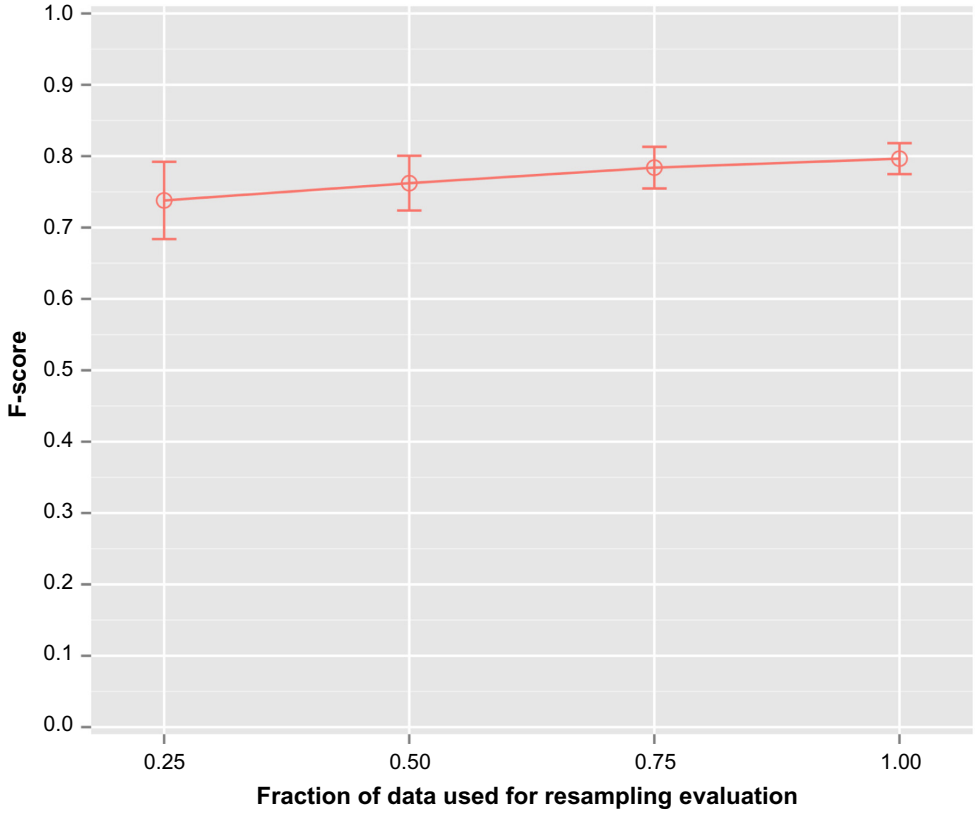

Figure 2. Classifier performance for different training data sizes.
Notes: Logistic regression models were trained on different sizes of data so as to observe their impact on the performance. Specifically, mean *F*-scores were calculated through resampling evaluation, where four different fractions of annotated data (0.25, 0.50, 0.75, and 1.00) were used to conduct the evaluation tests. The error bars represent one standard deviation from the mean.

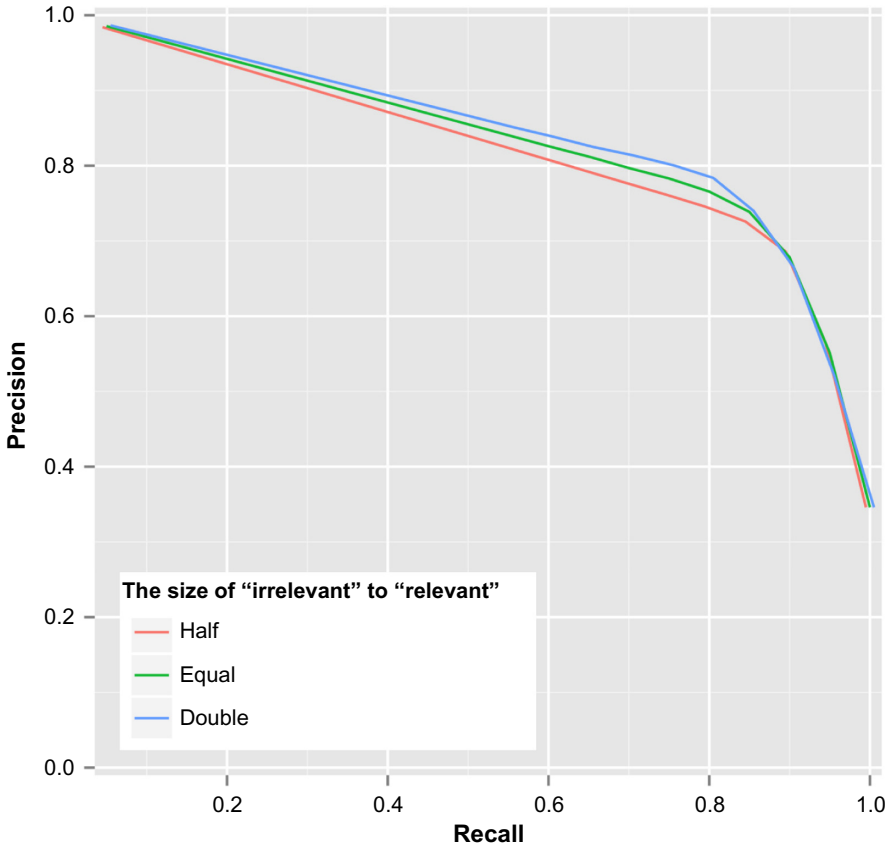

Figure 3. Classifier performance for different ratios of irrelevant to relevant instances.
Notes: Logistic regression models were trained on different ratios of positive and negative data so as to determine an appropriate ratio for the final model. In the original dataset, the size of irrelevant instances was double the size of relevant instances (Double). In addition to the original ratio, two different ratios, *Half* and *Equal*, were tested. Mean precisions and recalls were calculated through repeated resampling evaluation tests.

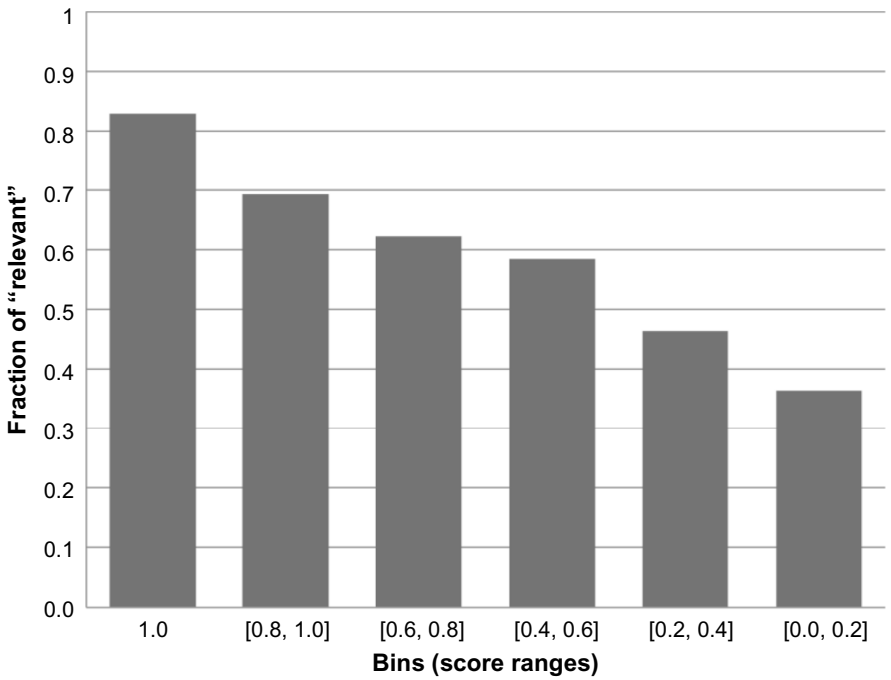

Figure 4. A fraction of relevant instances per bin (score range) on an unseen data set.
Notes: A logistic regression model was applied to a large collection of phrases in an unseen dataset, and phrases were sorted in bins according to prediction scores assigned by the classifier. One hundred phrases were sampled from each bin, and manually reviewed to estimate the fraction of relevant instances per bin.

Table 1. Frequent health issues in the customer reviews.

HEALTH ISSUES	UMLS CUI	PHRASE VARIANTS	FREQUENCY
Pain	C0030193	Pain(s), painful, hurt(s), hurting, ache	146
Diabetes	C0011849, C0375113	Diabetic(s), diabetes	122
Nausea	C0027497	Nausea, nauseous, nauseated, feeling sick	103
Headache	C0018681	Headache(s), head-ache, head pains	79
Morning sickness	C0240352, C0312416	Morning sickness	66
Upset stomach	C0235309	Upset stomach(s), upset tummy, stomach discomfort	42
Allergy	C0685900, C0700625, C0851444	Allergic, allergy, allergies	35
Diarrhea	C0011991	Diarrhea, loose stools	33
Acid reflux	C0017168	Acid reflux, gerd	23
Stomach ache	C0221512	Stomach ache(s), tummy ache(s), pain in the stomach	20

Note: The frequencies in the rightmost column were counted in the manually annotated data.

several broad categories as shown in Table 3. An *unclassified/unrelated to product* category was created to capture cases that were difficult to assign labels to. The percentage of each category is visualized in Figure 5.

Discussion

Health issues in the reviews. The most frequent health problems found in the grocery reviews (Table 1) are commonly encountered in our daily life. Given that many products are foods and drinks, a majority of the problems are symptoms related to the gastrointestinal system. Being the two explicit diseases in the list, diabetes and allergy both represent leading chronic conditions that concern people's diet decisions. As an example of significance, it was estimated that diabetes and prediabetes cost America $322 billion every year,[31] including direct medical expenditure and indirect loss of productivity. Based on the reviews, a good sign is that consumers do pay attention to prevention/management of diabetes in their shopping choices. On the other hand, there are still abundant opportunities to better integrate patient education into shopping applications especially for disease-specific population.

To give readers a feel on the products that frequently cause the health problems, in Table 2 we provide a couple of product types (instead of original maker/product names) for the top five health conditions of Table 1. Since we considered both positive and negative reviews as relevant, the products could be either beneficial or detrimental. A potential application along this line is to systematically collect/organize the peer-recommended products per health condition and share with the concerned communities. Note that this study focuses on grocery goods, and we expect that reviews on other product categories, such as sports, would reveal an even wider variety of health issues that involve injuries or ergonomics.

The categories we summarized in Table 3 confirm with general intuition on the issues consumers usually write about products. Adverse effect is probably the most serious that can be reported about a product, and it constitutes a substantial portion (~20%) among the health-related reviews. Although adverse reaction can be confounded by inappropriate use or an existing health problem, the reviews may serve as a valuable surveillance source and complement formally collected information. For example, the U.S. Department of Health and Human Services published a report[32] that pointed out

Table 2. Examples of product types per health issue.

PROBLEM	PRODUCT TYPE	EXAMPLES
Pain	Purified water	*There is no rational explanation–this water is amazing. [.] After a few hours the pain left.*
	Cherry tart juice	*I purchased this for my mother who was suffering from gout pain in her knee and big toe.*
Nausea	Vitamin supplements	*I purchased these drops in the hopes of helping my nausea due to morning sickness.*
	Ginger candy	*I didn't like the flavor and it actually make me more nauseous afterward.*
Headache	Caffeinated water	*Can be helpful for some allergies or headaches for some people too!*
	Energy drink	*I like the effects of this drink, however, it gives me a headache that I can't get rid of.*
Diabetes	Natural soda	*I would recommend it to anyone trying to cut out sugar or diabetics.*
	Baking mix	*I have a friend with diabetes and I am using it when he comes over to eat.*
Upset stomach	Ginger candy	*Great for upset stomachs*
	Sweetener	*I would get a horribly upset stomach within minutes of ingesting it.*

Table 3. Categories of health issues in the grocery reviews.

CATEGORY	EXAMPLES
Adverse effect	• *Major migraine trigger!* • *We don't have any issues with arthritis (yet) but find these caps creating finger aches and pains where there were none before.* • *Like several other posters, I started having severe bloating, cramping and diarrhea shortly after consumption that has now lasted for over 12 hours despite Gas-X and Pepcid.*
Benefit of product	• *It's supposed to help literally pull gingivitis out.* • *Peppermint is a well-known upset stomach soother and natural antispasmodic, so if you suffer from acid reflux or chronic heartburn, this tea is for you!* • *It handles the withdrawal effects from caffeine addiction while avoiding the obesity and diabetes causing effects of the corn syrup in the soda.*
First-person health problem	• *I adhere to a gluten free diet due to celiac disease.* • *Hi, i have arthritis in my legs and hope to ease the pain by drinking lemongrass tea with lemongrass oil (food grade).* • *Since I have osteoarthritis I was looking for something natural to help with flare ups.*
Third-person health problem	• *My daughter has asthma, eczema, and allergies, so we try to avoid as many of the "triggers" as possible.* • *My nephew-in-law is undergoing treatment for Lyme disease and his doctor wants him to go on a gluten free diet to help his body fight the disease.* • *My partner with type 2 diabetes started taking them for good health and its helped manage his blood sugar levels to where he is decreasing his daily insulin.*
Risk to health	• *Fructose is metabolized directly by the liver, which is different from other sugars, and can lead to fatty liver deposits.* • *This product or any product claiming to be sugar free have alcohol sugar in them, our small intestine is not able to absorb it.which cause bloating, diarrhea and flatulence!* • *If you eat normally these will clog your arteries, and contribute to your death of coronary artery disease.*
Unclassified/unrelated to product	• *Fatigue is unavoidable in long runs, but when added to that early morning non-caffeine sluggishness–things can get really ugly.* • *The good news is that I quit cigarettes in time to avoid most lung diseases.*

limitations of the Food and Drug Administration (FDA) adverse event reporting system for non-prescription dietary supplements. Since there is no premarket approval regulation, the FDA mostly relies on the consumer-initiated reporting. However, the mechanism is passive and not effective in tracking product information, prevalence, and trends. In contrast, reviews from major e-commerce sites have two advantages: (1) they are created actively by a large consumer population as part of regular business process, and (2) they cover diverse products and are usually already integrated with useful information about the products and consumers. Therefore, we believe automated syndromic surveillance over massive consumer reviews will benefit sensitivity, amass stronger signal, and avail richer attributes for analytic inquiry.

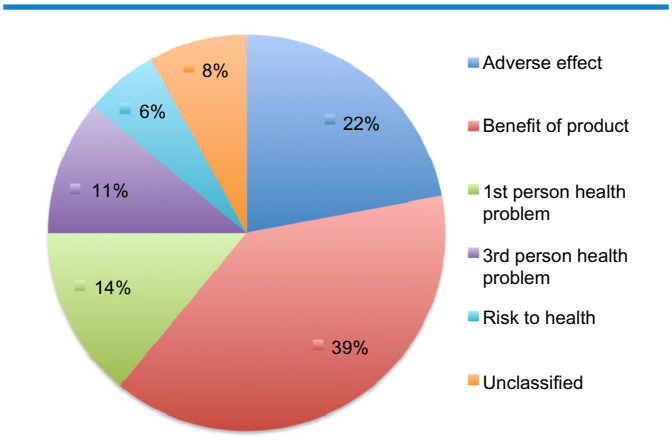

Figure 5. Proportions of the health issue categories.
Notes: One hundred concept phrases were sampled among those that were manually confirmed as relevant (health-related), and they were manually reviewed and categorized.

It is not surprising that a majority (~40%) of the health-related reviews comment about the beneficial effects of using the product. However, one caveat with those positive reviews is the difference between personally verified benefit versus benefit merely based on belief or layman knowledge passed among consumers. A closely related category is reviews that warn of certain risk to health, which in many cases may not be based on personal experience but reference to second-hand medical knowledge of unknown sources. The credibility of such health information and its influence over consumer decisions are interesting research topics to investigate.[33] The other two categories relate to existing health issues of the consumer her/himself or a close person the purchase was for. It is reasonable to see that such issues together constitute a substantial 25% of health-related reviews, and they justify the purchase decisions to resolve symptoms or avoid adverse effects given the underlying health condition. As discussed above, such problem-specific warning/recommendation may be systematically collected and shared among population with pertinent needs.

Challenges in mining health-related information from reviews. The data size can be tamed with advancement

of technology, but the intricate characteristics of the contents remain most challenging. During annotation of our training data, we debated over what should be defined as health-related (relevant). For example, when a review mentioned absence of side effects from a product, it could actually suggest the relative benefit of avoiding a health problem that concerns the consumer who had negative experience with using another symptom-triggering product. It can be difficult to differentiate between normal and abnormal findings. For instance, *thirsty* can be a temporary normal effect or a symptom or a more serious underlying condition. In the end, we decided on a definition that accepted many implicit cases as long as the consumer expressed concern about the product's effect on his/her or someone's health. Overall, the gain or loss of accepting/rejecting borderline cases was not significant, because the representative categories/patterns emerged as the data size grew large.

We observed considerable unexpected errors by the NLP engine, which had been tuned toward processing clinical documents. The general English of consumer reviews appear more diverse than the clinical sublanguage and can easily cause the engine to make mistakes. We summarized the common types of false extractions in Table 4. There are many ambiguous terms overlapping with medical usage, for example, idioms (*headache* used as metaphor) and abbreviations (*pat* as paroxysmal atrial tachycardia). A simple solution is to remove from the NLP engine the terms that are rarely used in the medical sense in consumer reviews, assuming laymen would rarely use terms used by professionals. Alternatively, the word sense disambiguation module in the NLP engine may be customized to the application domain, but that would require substantial efforts. Other types of errors include those introduced by false typo correction, which caused artificial ambiguity (eg, falsely correcting *stoke* into *stroke*). Since typo correction is still helpful given that there can be many misspellings in reviews, the desirable remedy should be expanding the lexicon of the NLP engine so as to avoid falsely triggering the typo correction. Additionally, we noted that health issues of nonhuman subjects (especially pets) were not uncommon in consumer reviews. In order to handle such variety of contents expected in product reviews, the human versus nonhuman challenge may deserve research for a dedicated solution.

Future Work

One of the major challenges noticed for the NLP engine was sense ambiguity as in Table 4. In our current study, we used bags of words features in the machine learning classifiers, but it would be of great interest to explore additional features, such as those based on word embedding. Apache Spark MLlib provides the implementation of word2vec, a word embedding technique, which computes distributed vector representation of words. Word embedding has been successfully employed in different NLP applications.[34–36] In terms of the classification categories, we did not differentiate customer reviews reporting positive effect of a product from those reporting negative one in the current study. We will seek developing a finer classifier that can predict the sentiment polarity. To detect weak signal of health conditions that are hypothesized with certain association or conflicting management concerns, we plan to mine co-occurring health issues mentioned in the reviews and inspect for any possible valid manifestation. Additionally, we are interested in applying our methods to other product

Table 4. False positives (FPs) of NLP extraction.

FP TYPE	EXPLANATION	EXAMPLE
Ambiguity	Idiomatic expression for complaint of trouble	*When I tried to read the nutrition facts, I got an instant headache while trying to make the words out.*
	"Worms" is used as a synonym of C0018889 Helminthiasis, a type of parasite infection	*I prefer to avoid artificial flavors and colors in my food, and these gummy worms are as satisfying as the more mainstream variety.*
	Patient is "dehydrated" versus the dehydration process in food industry	*This is a fantastic oat-based, crunchy granola with just the right amount of chunks of dark chocolate and dehydrated berries.*
	C0030587 Paroxysmal atrial tachycardia can be abbreviated as "pat"	*I add a pat of butter, a little maple syrup and some milk for a quick breakfast.*
Typo correction error	NLP considered "nostalgia" a typo of "notalgia", which is a synonym of C0004604 Back pain	*They evoke a nostalgia, and yet they are more complex and tastier than the one-note puffs of my youth.*
	NLP considered "stoke" as typo of "stroke", a synonym of C0038454 Cerebrovascular accident	*I stoke mine in freezer zip locks to keep them fresh in the freezer.*
Not on human	On cat	*Also, I took in a stray kitten who had an eye infection.*
	On dog	*My dog won't eat prescription dog food for her heart failure, so I'm winging it with people food- adult and baby.*
Semantic modifier	Analogy that refers to taste	*Personally I think all energy drinks taste pretty gross: either like carbonated cough syrup or something worse.*
	Part of organization name, not explicitly referring to problem	*The American Diabetes Association recommends that Sugar Alcohols not be consumed in excess of 20–50g per day.*

categories, such as sports/outdoors and beauty. Leveraging such knowledge harvested from massive and diverse customer reviews, a website summarizing the reported health issues can be created to guide consumers in their decision making, for example, creating a website analogous to ConsumerLab.com,[37] which summarizes test results of products related to health and nutrition.

Limitations

Although we set up and employed a scalable text analysis framework for our further exploration, we have not taken full advantage of the provided scalability in the reported work. In terms of the granularity of the target information, our study design of treating both positive and negative reviews as one single relevant class may not align with general interest, and our definition of the relevant cases was not free from subjectivity. Additionally, we did not perform systematic cleaning or reconciliation on the manual annotations. Due to limited resources, we did not perform evaluation of individual NLP modules, such as the phrase extractor and the concept searcher. Because our aim was to establish a general framework, we did not explore and customize features specific to the current data or thoroughly tune parameters of machine learning classifiers.

Conclusions

The existence of data recording many diverse aspects of our daily activities has opened new opportunities for informatics research on public health monitoring/promotion. In this study, we extracted health-related information from Amazon grocery product reviews by leveraging scalable analytic technologies. Benefiting from a big data scale-out framework, the NLP system completed processing 1.3 million reviews in 6 hours, despite several computationally expensive steps such as sentence parsing and concept searching. A random subset of about 5,000 disease/symptom phrases was manually annotated to train a logistic regression classifier based on Apache Spark MLlib. The high-confidence predictions of the classifier achieved a precision of 0.82. Such a classifier could be used to screen for health-related data in new consumer reviews. The health issues we found in the reviews are useful in terms of (1) complementing existing public health data sources, (2) empowering consumers for better-informed decisions, and (3) providing feedback to product manufacturers for improvement. On the technical side, our content and error analyses pointed out challenges in using NLP to process massive product reviews and in extracting health-related information from unconventional information sources. The study delivered useful insights for future research.

Acknowledgment

We thank our KPSC Medical Informatics team members Bayan Azima and Sunil Karumuri for their solid and timely technical support in preparing/processing the data.

Author Contributions

Conceived and proposed the research idea: JF. Designed and implemented the experiments: MT, SST. Analyzed the data: MT, SST, SD, JF. Wrote the first draft of the manuscript: JF. Contributed to the writing of the manuscript: MT, SST, SD, DSZ, JF. Agreed with manuscript results and conclusions: MT, SST, SD, DSZ, JF. Jointly developed the structure and arguments for the paper: MT, SST, SD, JF. Made critical revisions and approved the final version: MT, SST, SD, DSZ, JF. All the authors reviewed and approved the final manuscript.

REFERENCES

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128–44.
2. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*. 2011;18(5):540–3.
3. Corley CD, Cook DJ, Mikler AR, Singh KP. Text and structural data mining of influenza mentions in web and social media. *Int J Environ Res Public Health*. 2010;7(2):596–615.
4. Ofoghi B, Mann M, Verspoor K. Towards early discovery of salient health threats: a social media emotion classification technique. In: Pacific Symposium on Biocomputing. Kohala Coast, Hawaii: World Scientific; Vol 21; 2016:504–15.
5. Aphinyanaphongs Y, Lulejian A, Brown DP, Bonneau R, Krebs P. Text classification for automatic detection of E-cigarette use and use for smoking cessation from twitter: a feasibility pilot. In: Pacific Symposium on Biocomputing. Kohala Coast, Hawaii: World Scientific; Vol 21; 2016:480–91.
6. Sarker A, Ginn R, Nikfarjam A, et al. Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform*. 2015;54:202–12.
7. Lindner M. Global e-commerce sales set to grow 25% in 2015; 2015. Available at: https://www.internetretailer.com/2015/07/29/global-e-commerce-set-grow-25-2015. Accessed February 3, 2016.
8. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat Lang Eng*. 2004;10(3–4):327–48.
9. Zaharia M, Chowdhury M, Das T, Dave A. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: NSDI'12 Proc 9th USENIX Conf Networked Syst Des Implement. San Jose, CA: USENIX Association; 2012:2.
10. Feiguina O, Lapalme G. Query-based summarization of customer reviews. *Adv Artif Intell*. 2007;4509:452–63.
11. Tang H, Tan S, Cheng X. A survey on sentiment detection of reviews. *Expert Syst Appl*. 2009;36(7):10760–73.
12. Korfiatis N, García-Bariocanal E, Sánchez-Alonso S. Evaluating content quality and helpfulness of online product reviews: the interplay of review helpfulness vs. review content. *Electron Commer Res Appl*. 2012;11(3):205–17.
13. Sullivan R, Sarker A, O'Connor K, Goodin A, Karlsrud M, Gonzalez G. Finding potentially unsafe nutritional supplements from user reviews with topic modeling. In: Pacific Symposium on Biocomputing. Kohala Coast, Hawaii: World Scientific; Vol 21; 2016:528–39.
14. Schneeweiss S. Learning from big health care data. *N Engl J Med*. 2014; 370:2161–3.
15. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Heal Inf Sci Syst*. 2014;2:3.
16. Nodarakis N, Sioutas S, Tsakalidis A, Tzimas G. Large scale sentiment analysis on twitter with spark. In: 1st International Workshop on Multi-Engine Data AnaLytics (EDBT/ICDT Workshops). Bordeaux, France: CEUR-WS.org; 2016.
17. Statistics and facts about Amazon. Available at: http://www.statista.com/topics/846/amazon/. Accessed February 6, 2016.
18. McAuley J, Pandey R, Leskovec J. Inferring networks of substitutable and complementary products. In: Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'15). Sydney, NSW, Australia: Association for Computing Machinery; 2015:12.
19. JSON – JavaScript Object Notation. Available at: http://www.json.org/. Accessed February 6, 2016.
20. Torii M, Fan J, Yang W, et al. Risk factor detection for heart disease by applying text analytics in electronic medical records. *J Biomed Inform*. 2015;58(suppl):S164–70.

21. Apache ActiveMQ. The Apache Software Foundation. Available at: http://activemq.apache.org/. Accessed February 3, 2016.
22. Apache Hadoop. The Apache Software Foundation. Available at: http://hadoop.apache.org/. Accessed February 6, 2016.
23. Apache Pig. The Apache Software Foundation. Available at: http://pig.apache.org/. Accessed February 6, 2016.
24. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: Proc. of the OSDI – Symp. on Operating Systems Design and Implementation. San Francisco, CA: USENIX Association; 2004:137–49.
25. Apache Spark. The Apache Software Foundation. Available at: http://spark.apache.org/. Accessed February 6, 2016.
26. Apache Lucene. The Apache Software Foundation. Available at: https://lucene.apache.org/. Accessed February 6, 2016.
27. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform*. 2001;84(pt 1):216–20.
28. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press; 2008.
29. Zhang T, Oles FJ. Text categorization based on regularized linear classification methods. *Inf Retr Boston*. 2001;4(1):5–31.
30. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011;18(5):552–6.
31. American Diabetes Association. The cost of diabetes; 2015. Available at: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html. Accessed February 7, 2016.
32. Office of Inspector General, Department of Health and Human Services, United States. Adverse event reporting for dietary supplements. An inadequate safety valve, April 2001, OEI-01-00-00180. Available at: http://oig.hhs.gov/oei/reports/oei-01-00-00180.pdf. Accessed May 27, 2016.
33. Saurí R, Pustejovsky J. Are you sure that this happened? Assessing the factuality degree of events in text. *Comput Linguist*. 2012;38(2):261–99.
34. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc*. 2015;22(3):671–81.
35. Wu Y, Xu J, Zhang Y, Xu H. Clinical abbreviation disambiguation using neural word embeddings. In: Proceedings of BioNLP 15. Beijing, China: Association for Computational Linguistics. Beijing, China: Association for Computational Linguistics; 2015:171–6.
36. Zhang S, Grave E, Sklar E, Elhadad N. Longitudinal Analysis of Discussion Topics in an Online Breast Cancer Community using Convolutional Neural Networks; 2016. Available at: http://arxiv.org/abs/1603.08458. Accessed April 27, 2016.
37. ConsumerLab.com – independent tests and reviews of vitamin, mineral, and herbal supplements. Available at: http://www.consumerlab.com/. Accessed April 6, 2016.