FINDING POTENTIALLY UNSAFE NUTRITIONAL SUPPLEMENTS FROM USER REVIEWS WITH TOPIC MODELING

RYAN SULLIVAN*, ABEED SARKER, KAREN O'CONNOR, AMANDA GOODIN, MARK KARLSRUD and GRACIELA GONZALEZ

Department of Biomedical Informatics, Arizona State University,
Scottsdale, AZ 85259, USA
**E-mail: rpsulli@asu.edu, abeed.sarker@asu.edu, karen.oconnor@asu.edu, agoodin@asu.edu, mkarlsru@asu.edu, graciela.gonzalez@asu.edu*

Although dietary supplements are widely used and generally are considered safe, some supplements have been identified as causative agents for adverse reactions, some of which may even be fatal. The Food and Drug Administration (FDA) is responsible for monitoring supplements and ensuring that supplements are safe. However, current surveillance protocols are not always effective. Leveraging user-generated textual data, in the form of Amazon.com reviews for nutritional supplements, we use natural language processing techniques to develop a system for the monitoring of dietary supplements. We use topic modeling techniques, specifically a variation of Latent Dirichlet Allocation (LDA), and background knowledge in the form of an adverse reaction dictionary to score products based on their potential danger to the public. Our approach generates topics that semantically capture adverse reactions from a document set consisting of reviews posted by users of specific products, and based on these topics, we propose a scoring mechanism to categorize products as "high potential danger", "average potential danger" and "low potential danger." We evaluate our system by comparing the system categorization with human annotators, and we find that the our system agrees with the annotators 69.4% of the time. With these results, we demonstrate that our methods show promise and that our system represents a proof of concept as a viable low-cost, active approach for dietary supplement monitoring.

Keywords: Dietary Supplements, Pharmacovigilance, Natural Language Processing, Latent Dirichlet Allocation, Public Health Surveillance, Social Media Mining.

1. Introduction

According to the Dietary Supplement and Health Education Act (DSHEA),[1] dietary supplements (often referred to as nutritional products) are intended to supplement diet, intended for oral use, contain one or more dietary ingredients or their constituents, and are labeled on the packaging as dietary supplements. 50% to 70% of the general population in the United States uses a dietary supplement either for their purported benefits in maintaining good health or for the treatment of various diseases.[2–5] Evidence from multiple surveys suggests that dietary supplement users are more likely than non-users to adopt a number of positive health-related habits.[6] Thus, dietary supplements have become an integral part of health and wellness, and many health professionals and dietitians use and recommend their use.[4]

Despite the usefulness of dietary supplements, their widespread usage, and the perception that they are safe for use, they have been identified as causative agents for a variety of adverse

This work was funded (in part) by the University of Maryland, Joint Institute for Food Safety and Applied Nutrition through a cooperative agreement with the U.S. Food and Drug Administration (or FDA), #FDU001418.

reactions. For example, consumption of Chinese herbs that contain aristolochic acid (Mu Tong) has been reported to be associated with an increased risk of urinary tract cancer,[7] and more recently, the product OxyElite Pro® was recalled by the U.S. Food and Drug Administration (FDA) in November 2013[a] after possible links between the product and both liver failure and non-viral hepatitis were discovered.

Currently in the United States, the FDA regulates both finished dietary supplement products and dietary ingredients under a different set of regulations than those covering conventional food and drug products (prescription and over-the-counter).[8] Under the DSHEA[1] a manufacturer is responsible for ensuring that a dietary supplement or ingredient is safe before it is marketed. The FDA is responsible for taking action against any unsafe dietary supplement product after it reaches the market, and intervening if there is misleading product information. Generally, manufacturers do not need to register their products with the FDA nor do they need to get FDA approval before producing or selling dietary supplements. The responsibility of the manufacturer is to ensure that product label information is truthful and not misleading, that the product complies with the Dietary Supplement Current Good Manufacturing Practices (cGMPS) for quality control, and to submit to the FDA all serious adverse events[b] reports associated with use of the dietary supplement in the United States.

Under current adverse event monitoring protocols drug manufacturers and consumers can report adverse events caused or suspected to be caused by a dietary supplement using the Safety Reporting Portal.[c] Safety reports can be voluntarily submitted by manufacturers, packers, holders, researchers, or end users. However, numerous pharmacovigilance studies have revealed the ineffectiveness of self-reporting systems,[9] with some studies showing that only about 10% of adverse reactions generally reported.[10] There are many possible reasons for the low reporting numbers; a manufacturer may be reluctant to admit fault, or users may not report events (particularly for non-lethal events) to the manufacturer or even health care providers. Furthermore, even when a consumer has a serious event and goes to a poison center and a report is created, the FDA may not receive it. A 2013 Government Accountability Office report found that from 2008 through 2010 poison centers received over 1000 more reports than the FDA.[11] These facts clearly demonstrate that active surveillance is essential to the FDA's public health mandate with respect to dietary supplements. Although alternative sources (such as user comments from health forums or tweets) have been shown as potential sources for monitoring adverse reactions associated with prescription drugs,[12] there is still a research gap on active monitoring of dietary supplements.

[a] `http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm374742.htm`,
`http://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm370849.htm`.
Accessed: 7/10/2015.

[b] A serious adverse event is defined by the FDA as any adverse dietary supplement experience occurring at any dose that results in any of the following outcomes: death, a life-threatening adverse dietary supplement experience, inpatient hospitalization or prolongation of existing hospitalization, a persistent or significant disability/incapacity, or a congenital anomaly/birth defect.
(`https://www.federalregister.gov/articles/2007/06/25/07-3039/`
`current-good-manufacturing-practice-in-manufacturing-packaging-labeling-or-holding-operations-for#`
`h-493`. Accessed: 7/29/2015.)

[c] `https://www.safetyreporting.hhs.gov/fpsr/WorkflowLoginIO.aspx`. Accessed 7/10/2015.

Due to the strong motivation for active, low-cost monitoring systems for dietary supplements, we focused our study on extracting signals indicating the safety of dietary supplements from publicly available data on the Internet. In particular, we collected and automatically processed a large set of Amazon.com reviews, and used that information to predict the safety of the products. Our approach generates topics for each dietary supplement product based on its reviews, and uses these topics, with the assumption that the topics capture the semantic concepts associated with adverse effects, to rank the relative safety of individual products as compared to others in the same product class.

To generate the topics, we use a fully unsupervised variant of Latent Dirichlet Allocation (LDA).[13] Our approach biases the topic model by guaranteeing that tokens that match adverse reactions, based on ADRs listed in the SIDER database[d], will be limited to a sub-set of topics, and uses the topic distribution of a given product's reviews to score and rank that product. Essentially, the topic distributions are used as weights to score each product based on how much of the texts in its reviews appeared to be *generated* by the adverse reaction topics.

We consider three categories for each product: "high potential danger", "average potential danger" and "low potential danger," and compare the predictions of our system to a small set of 18 products categorized by human annotators. We find that our system agrees with the human rankings 69.4% of the time. Figure 1 visually illustrates our pipeline. We discuss the different components of the pipeline in the following sections, commencing with an overview of related literature.

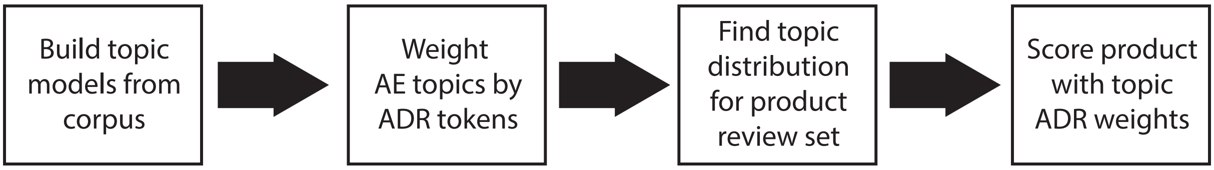
Build topic models from corpus
Weight AE topics by ADR tokens
Find topic distribution for product review set
Score product with topic ADR weights

Fig. 1. System pipeline.

2. Related work

For public health issues, mining user-generated content has been shown to be a valuable resource of information, particularly because of the large volume and the possibility of real-time analysis.[14–16] Due to the underutilization of traditional reporting avenues,[17] detecting prescription drug ADR mentions in social media posts is an area that has seen a flurry of recent research. Leaman *et al.*[12] performed some of the earliest research in this area, using data from DailyStrength[e] to determine the feasibility of using lexicons for finding and extracting ADRs from user comments. Subsequent research was performed by Benton *et al.*[18] and Yates and Goarian,[19] and also used keyword based approaches, supplemented by synonym sets of lay vocabulary, to identify drug adverse events from social media sites.

[d] http://sideeffects.embl.de/
[e] http://www.dailystrength.org/

Current research in this space has also utilized NLP and ML techniques to overcome the shortcomings of lexicon-based approaches. For example, Nikfarjam and Gonzalez[20] and Yang *et al.*[21] both use association rule mining for ADR-related tasks using user-generated health text. Supervised text classification approaches have also been popular, particularly the use of Support Vector Machines (SVMs) (*e.g.*, Bian *et al.*,[22] Sarker and Gonzalez[23]).

Recent research has also seen the application of unsupervised approaches. For example, the study by Yang *et al.*[24] showed that LDA can be combined with a partially supervised classification approaches to create a classifier to locate consumer ADR messages in on-line discussion groups, and a study by Li *et al.*[25] showed that adding topics generated by LDA as a feature for an assertion classifier lead to a significant improvement in classification. Furthermore, Bisgin *et al.*[26] demonstrated that topics generated by LDA using drug labels as documents could be used to group drugs in a statistically significant way, which could be useful for discovering new relationships between drugs.

3. Methods

Our approach involves learning a probabilistic topic model that is partially based on background knowledge in the form of a dictionary of adverse reactions. We then build a weight map for our topic model where each topic is mapped to a value estimating how much each topic represents the ADRs. Finally, we use our topic model and our weight map to assign a single score to each product indicating the extent to which the reviews can be attributed to adverse reactions.

3.1. *Data*

Using a web crawler, we created a corpus of approximately 40,000 Amazon.com reviews from 2708 products[f]. The products chosen were those categorized by Amazon.com as “Herbal Supplements,” “Sports Nutrition,” “Supplements” and “Weight Loss.” Our corpus consists of all products and from all their subcategories. Sample reviews for two products are shown in Table 1. These examples are representative of what is found across the review corpus and present examples of adverse reactions and indications. Furthermore, these examples show the varying seriousness of adverse reactions within the reviews and also give an example of a reviewer talking about a AE, as opposed to mentioning the event.

3.2. *LDA using background knowledge*

Our approach is driven by a variant of LDA.[13] LDA is an unsupervised technique, generally used for topic modeling, which builds a generative model for the data. Generative models are models which, given some parameters, could have randomly generated the observed data. In our specific case, we attempt to estimate the document-topic distributions and the topic-token distributions from which it would be possible to generate our text corpus. A document is generated by an LDA model one token at a time. The process begins by sampling the per-document topic distribution to choose a topic and then sampling the token distribution for

[f] Reviews were captured on 5th March 2014 from `http://www.amazon.com/b?node=37644410`

Table 1. Sample user reviews for two dietary supplement products.

Product: **batch5 Extreme Thermogenic Fat Burner**	Product: **NOW Foods Bromelain**
This pills dont work at all. Its just another pill with to much caffeine and makes you cranky, edgy and nervous.	This is just fine.....not sure what it was for. I do believe it is helping with my sinus problems, at least I haven't had any lately.
I take this product before i work out and i feel more energetic and i get a feeling of well being and it last long after im done working out. I definitely recommend B4.	the product caused adverse reactions for me and could not tolerate, had back pain and right right kidney pain and decreased urine output was not good for me.
I felt awful after I took it got a terrible niacin rush would never take it again side effects are scary	This product has helped me with the pain I have in my joints due to arthritis. My knees and hands were so bad before, but after just a couple of weeks I have gotten amazing relief.

the chosen topic to pick a word. The chosen word is added to document, and the process is repeated for the length of the document.

Our process is a variant of LDA which seeks to take advantage of background knowledge, which in our case is a dictionary of adverse reactions. Our intent is to generate topics that are semantically similar to the adverse reactions. We accomplish our goal by developing an LDA variant which uses a second per-document topic Dirichlet distribution ($Dirichlet(\alpha')$), which when sampled, will return a multinomial distribution over a sub-set of topics. This distribution over a subset of topics is then sampled to generate words that are known to be from our dictionary. This variant can be thought of as two parallel instances of topic modeling. One instance consisting of the tokens found in the dictionary and encompassing a subset of topics, and a second instance of the standard LDA for all topics and all non-ADR tokens.

Formally, our approach can be described as follows:

Let D be a collection of documents. For each $d \in D$ of length N, let $f_d : \{1, \ldots, N\} \to \{0, 1\}$ be an indicator function that maps an index in d to 1 when the word at the index is part of the background knowledge. To generate the collection D:

(1) For each topic k, draw a multinomial token distribution ϕ_k from $Dirchlet(\beta)$
(2) For each Document $d \in D$:
  (a) Draw a multinomial topic mixture θ from $Dirchlet(\alpha)$
  (b) Draw a multinomial topic mixture θ_{sub} from $Dirchlet(\alpha')$
  (c) Choose a document length N
  (d) For each token $0 \leq i < N$ in document d
    i. if $f_d(i) = 1$ choose topic z_i from θ_{sub}, else choose topic z_i from θ
    ii. Choose word w_i from ϕ_{z_i}

Figure 2 presents the plate notation for this variation of LDA.

This method is based on the general assumption that tokens which match the entries in

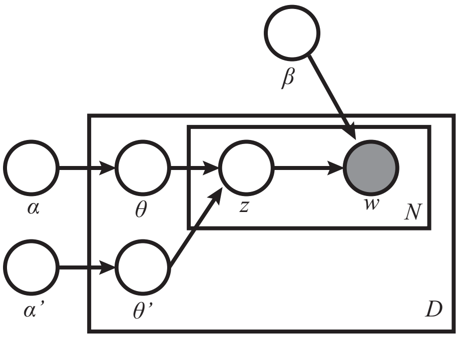

Fig. 2. Plate notation of our LDA model

the ADR dictionary could only have been chosen from a marked subset of topics. Though we do not label topics, we guarantee that tokens that match the tokens in the ADR dictionary are restricted to those subsets. The intent of this restriction is that the subset of topics containing ADR tokens will also contain tokens that are semantically similar to ADRs, but do not appear in the ADR dictionary.

Our approach was developed as an extension of the ParallelTopicModel class within the Mallet machine learning toolkit.[27] The ParallelTopicModel class is a implementation of the algorithm presented by Newman, et al.[28] and can be viewed as an approximation to Gibbs-sampled LDA. Our pre-processing consisted of removing stop words, and representing every instance of multi-token dictionary ADRs in the text as a single token. We chose to use 100 topics, and chose a subset size of 10. For our priors we chose standard values, $\alpha = 0.1$, $\alpha' = 0.1$ and $\beta = 0.01$. To learn our model we chose to use 10000 iterations of Gibbs Sampling and use a burn in of 1000 iterations. Table 2 provides examples of the top 15 tokens from selected topics from the category "SportsNutrition/Thermogenics/Fat Burners." The ADR topics are in bold.

3.3. *Scoring products based on topics*

Our system uses the Topic Models of the review set to generate a score for each product. Each topic is a distribution of tokens, so every token within a topic carries a weight as to how important that token is within its topic. We sum the weights of our known ADR tokens within each topic, and for each topic create a topic ADR weight. These topic ADR weights are the primary component of our scoring system.

To score each product, we first represent the product as a single document containing all the reviews. We then use the Mallet Topic Inferencer to estimate the distribution of topics for the product reviews. This provides us with information about how much of the review text was likely 'generated' by each topic, or what percent of the reviews can be explained by each given topic. We combine the topic percentages with the per-topic ADR weights to score each product and then normalize the product scores across all products within a category. An example of our scoring can be seen in table 3.

We choose to score products with respect to their Amazon category. That is, as opposed to building a topic model based on the full corpus, we build topic models for each category

Table 2. Tokens from Selected Topics of 'Fat Burners':

Topic 0	stomach, gas, doesn, problems, issues, give, product, upset, don, bloating, system, digestive, easy, bad, products
Topic 1	energy, boost, give, feel, extra, day, product, workout, focus, gave, jitters, work, workouts, felt, level, feeling, caffeine
Topic 3	blood, sugar, levels, body, cancer, health, diabetic, problems, insulin, liver, people, research, due, level, heart
Topic 47	oil, punch, meat, red, chicken, fish, fruit, eat, eggs, veggies, fruits, eating, vegetables, tropical, vegetable
Topic 75	lost, pounds, lbs, ve, weeks, months, weight, week, lose, taking, month, days, gained, started, pound
Topic 98	free, gluten, lactose, soy, intolerant, dairy, organic, milk, grass, fed, cows, wheat, product, gmo, products

we wish to evaluate. We also only score products in relation to other products in the same category. This was done because when the full corpus is used to generate topic models, we found that when one product has a strong co-occurrence with one type of ADR, the topics related to that ADR became more of a topic for the product class. In those cases, the ADR topics would represent the products with those adverse events, and not the adverse events within the product reviews. We also chose to only score the products that had at least 25 reviews because products with a low number of reviews do not have enough text for scoring to be accurate.

Table 3. ADR Score for product: Dexatrim Max Comple-7

Topic	Topic ADR Weight	ADR examples from topic	Topic Percent	Topic ADR weight
Topic 0	30	birth_defects(6.0), chest_pains(4.0)	0.01378	0.413
Topic 1	139	bloating(11.0), diarrhea(7.0)	0.00182	0.252
Topic 2	111	gas(14.0), headaches(12.0)	0.01138	1.263
Topic 3	41	liver_damage(6.0), loss_of_weight(4.0)	4.635 E-4	0.019
Topic 4	522	jittery(72.0), headache(67.0)	0.03975	20.749
Topic 5	202	gain_weight(18.0), feel_sick(16.0)	0.01276	2.577
Topic 6	131	jittery(46.0), heart_attack(12.0)	0.00283	0.370
Topic 7	53	hunger_pains(5.0), reduced_appetite(5.0)	0.01322	0.700
Topic 8	91	inflammation(10.0), joint_pain(7.0)	7.46 E-6	6.78 E-4
Topic 9	150	palpitations(21.0), high_blood_pressure(11.0)	0.016450	2.46
			ADR Score: 28.803	

4. Evaluation and results

Our primary goal is to develop a system to help identify potentially dangerous nutritional supplements. The majority of our evaluation is related to that primary goal. However, because such a large portion of our work is based on our variant on LDA, we feel it is necessary to provide an evaluation of that aspect of our methodology.

4.1. *Validation of background knowledge driven LDA*

To validate our methodology, we used the Twitter Adverse Drug Reaction corpus from Ginn *et al.*[29] and compared the ADR scores of the tweets annotated with adverse reactions to those tweets with no ADRs. We compared the ADR scores generated with topics from our variant to the scores generated with topics from standard LDA. We found that with our variant, tweets with an annotated adverse reaction on average had a ADR score 1.89 times bigger than the score of tweets without any adverse reactions. This can be compared to standard LDA, where the ADR tweets had a score on average of 1.56 times bigger than the non-ADR tweets. We also compared the tokens within the topics for both standard LDA and our variation. We examined the correlation between the weight of tokens from the SIDER database and tokens annotated as ADRs (not in the database). We found that the R-squared value for the correlation between known ADR tokens and annotated ADR tokens within topics was 0.356 for normal LDA and 0.445 for our variation. These results provide evidence that our variant on LDA does create topics which better capture adverse drug reactions.

4.2. *Evaluation of 'ADR Score'*

We evaluated our ADR Score results by having human annotators categorize products from within a category, and then comparing the categorization to our rankings. We chose to use the categories of "SportsNutrition/Thermogenics/Fat Burners" and "WeightLoss/AppetiteControl&Suppressants" for evaluation. From those categories we chose 9 random products, three from the top third, three from the middle, and three from the bottom third of the list of products within the category ranked by ADR score. Two human annotators then categorized each product, and we compared our automatically generated categorization to the annotator categorization.

4.2.1. *Human categorization of products*

The user comments for nine products from the "SportsNutrition/Thermogenics/Fat Burners" class and 'WeightLoss/AppetiteControl&Suppressants" class were manually reviewed by two expert annotators to assess the results of the classifier. For each product, the annotator classified the product as having either a high, average or low potential for ADRs. Each annotator assessed the ADR potential by a variety of indicators, including: comparing the number of comments with ADR mentions from the number of comments overall; the severity of the ADR mentioned; and the potential for adverse reactions from the ingredients in the supplement.

4.3. *Results*

Table 4 and Table 5 present the comparison of human annotated classification to the classification based on the 'ADR Score' for the class "SportsNutrition/Thermogenics/Fat Burners' and the class "WeightLoss/AppetiteControl&Suppressants." Over these two categories, The annotator agreement was 61.1 %. The system accuracy with respect to Annotator 1 is 66.6% and the accuracy with respect to Annotator 2 is 72.2%, and the average accuracy of our system is 69.4% over the two categories.

Table 4. Comparison of annotator categorizations with our systems categorizations for SportsNutrition/Thermogenics/Fat Burners.

Product	Human Annotator 1	Human Annotator 2	ADR Score	ADR Score Category
batch5 Extreme Thermogenic Fat Burner	High Potential	Average Potential	0.336	Average Potential
BPI Sports B4 Fat Burner	High Potential	High Potential	1.0	High Potential
Buy Garcinia Cambogia Extract With Confidence	Low Potential	Low Potential	0.129	Low Potential
Cellucor D4 Thermal Shock Thermogenic Fat Burner	High Potential	High Potential	0.614	High Potential
Garcinia Cambogia Drops	Low Potential	Low Potential	0.120	Low Potential
Liporidex MAX w Green Coffee Ultra	Average Potential	Average Potential	0.371	Average Potential
Raspberry Ketones The ONLY 250 mg PURE Raspberry Ketone Liquid	Low Potential	Low Potential	0.186	Low Potential
SafSlim Tangerine Cream Fusion	Low Potential	Average Potential	0.341	Average Potential
VPX Meltdown	Average Potential	High Potential	0.685	High Potential

5. Discussions and future work

The primary goal of this work is to use unsupervised NLP techniques for low-cost, active monitoring of dietary supplements, and with our results we have presented a promising proof-of-concept system. This system has shown to be reasonably accurate in identifying products with above-average potential for adverse reactions, especially when the results are considered with respect to the annotator agreement.

Through the process of error analysis, we found three important potential limitations of our system: The system treats all adverse reactions equally, it treats ADRs and indications equally, and it cannot differentiate real and fake reviews. In dietary supplement monitoring, a single serious adverse effect is given significantly more weight than multiple non-serious reactions. Currently, our system has no way to weigh the reactions, and thus numerous trivial reactions will generate a higher score than one serious adverse reaction. This particular case did lead to a disagreement between the annotators and the system, where one annotator found a product to have a higher potential than our system due to a small number of serious adverse reactions.

Table 5. Comparison of annotator categorizations with our systems categorizations for WeightLoss/AppetiteControl&Suppressants.

Product	Human Annotator 1	Human Annotator 2	ADR Score	ADR Score Category
Nature's Way Metabolic ReSet	Low Potential	Average Potential	0.385	High Potential
Burn + Control Weight-loss Gourmet Instant Coffee by Javita	Low Potential	Low Potential	0.058	Low Potential
Garcinia Cambogia Extract Pure (60% HCA)	Low Potential	Low Potential	0.0527	Low Potential
Garcinia Cambogia Extract Pure Premium Ultra	Low Potential	Low Potential	0.129	Average Potential
Life Extension Decaffeinated Mega Green Tea Extract	High Potential	Low Potential	0.366	High Potential
Garcinia Cambogia Liquid Weight Loss Diet Drops	Low Potential	Low Potential	0.0	Low Potential
LipoBlast Extreme Diet Pills/Energy Boosters/Appetite Suppressant	High Potential	Average Potential	0.128	Average Potential
MetaboLife Ultra, Stage 1	High Potential	Low Potential	0.335	High Potential
Saffron Extract - Appetite Suppressant	Average Potential	Low Potential	0.125	Average Potential

Indications can be defined as the reason why a consumer is taking a drug or supplement, and in many cases, indication tokens are adverse reaction tokens. The primary difference between indications and ADRs is how the reaction relates to the user with respect to the drug. For example, the ADR tokens in the phrase "*'This product has helped me with the pain I have in my joints due to arthritis*" are very similar to the ADR tokens in the phrase "*the product caused adverse reactions for me and could not tolerate, had back pain and right kidney pain and decreased urine output*," yet they are very different semantically. As the system currently works, a product that has many indications will be scored similarly to one with many adverse reactions, as the current system does not take into account the semantic relationship between a potential ADR term and the rest of the sentence.

Finally, the system is currently unable to identify fake reviews. Dietary supplement manufacturers are known to provide free products to those who write reviews, and for some products we found that there were a non-trivial amount of fake positive reviews. Because we are examining the percentage of the reviews that is generated by the ADR topics, these fake reviews affect our ranking.

The monitoring of dietary supplements is a challenging task due to both the sheer number of supplements on the market and the limited man-power of the FDA. Despite the current limitations, our system produces very promising results. In particular, this system shows the validity of an unsupervised NLP approach for this task, while also serving as a promising proof-of-concept system.

The current limitations also provide a roadmap for future work. We plan on exploring other variations of LDA such as the Topic Aspect Model[30] and multi-grain topic models,[31] to incorporate aspects of those approaches into our work. We feel these techniques are promising solutions that can help distinguish adverse reactions from indications.

We also plan on incorporating the work presented in Leaman et al.[12] to add ADR named entity recognition to our pipeline. This will allow our system to use more then just a dictionary of known adverse reaction tokens when learning the ADR topics. Furthermore, we plan on adding 'fake rule detection' to the pipeline, following the work of Lau et al.[32] In addition, we plan on expanding the system to include text from other sources, including Internet message boards. There are very active on-line communities which discuss nutritional supplements, and this textual data would add to our corpus and help increase the accuracy of our categorization. Finally, we plan to expand our evaluation and include a larger variety of product categories.

Our experiments show that large amounts of user-generated data, which is readily available, may be used to automatically identify high-risk dietary supplements. The identified supplements can then be marked for further investigation by the Center for Food Safety and Applied Nutrition (CFSAN). We hypothesize that this unsupervised NLP technique will provide valuable early signals of suspected associations between CFSAN-regulated products and adverse reactions. Based on our promising results, we envision that this technique will act as a crucial source for safety signals associated with dietary supplements and may eventually provide the ability to detect problematic supplements earlier and more cost effectively than current methods.

References

1. U. S. Food and Drug Administration, Dietary Supplement Health and Education Act of 1994 `http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/SignificantAmendmentstotheFDCAct/ucm148003.htm`, (1994).
2. K. Radimer, B. Bindewald, J. Hughes, B. Ervin, C. Swanson and M. F. Picciano, *American Journal of Epidemiology* **160**, 339 (2004).
3. B. B. Timbo, M. P. Ross, P. V. Mccarthy and C. T. J. Lin, *Journal of the American Dietetic Association* **106**, 1966 (2006).
4. A. Dickinson, L. Bonci, N. Boyon and J. C. Franco, *Nutrition Journal* **11** (2012).
5. K. Fisher, R. Vuppalanchi and R. Saxena, *Archives of Pathology and Laboratory Medicine* **139**, 876 (2015).
6. A. Dickinson and D. MacKay, *Nutrition Journal* **13** (2014).
7. M.-N. Lai, S.-M. Wang, P.-C. Chen, Y.-Y. Chen and J.-D. Wang, *Journal of the National Cancer Institute* **102**, 179 (2010).
8. U. S. Food and Drug Administration , Dietary Supplements `http://www.fda.gov/Food/DietarySupplements/` (April, 2015).
9. A. Sarker, A. Nikfarjam, K. OConnor, R. Ginn, G. Gonzalez, T. Upadhaya, S. Jayaraman and K. Smith, *Journal of Biomedical Informatics* **54**, 202 (February 2015).
10. R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan and C. Friedman, *Clinical pharmacology and therapeutics* **91**, 1010 (2012).
11. U. S. G. A. Office, *DIETARY SUPPLEMENTS: FDA May Have Opportunities to Expand Its Use of Reported Health Problems to Oversee Products*, tech. rep., United States Government Accountability Office (03 2013).
12. R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang and G. Gonzalez, Towards internet-

age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks, in *Proceedings of the 2010 workshop on biomedical natural language processing*, 2010.

13. D. M. Blei, A. Y. Ng and M. I. Jordan, *the Journal of machine Learning research* **3**, 993 (2003).
14. M. J. Paul and M. Dredze, You are what you tweet: Analyzing twitter for public health., in *ICWSM*, 2011.
15. M. Szomszor, P. Kostkova and E. De Quincey, # swineflu: Twitter predicts swine flu outbreak in 2009, in *Electronic Healthcare*, (Springer, 2012) pp. 18–26.
16. Y. T. Yang, M. Horneffer and N. DiLisio, *Journal of public health research* **2**, p. 17 (2013).
17. L. Hazell and S. A. Shakir, *Drug Safety* **29**, 385 (2006).
18. A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard and J. H. Holmes, *Journal of biomedical informatics* **44**, 989 (2011).
19. A. Yates and N. Goharian, Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites, in *Advances in Information Retrieval*, (Springer, 2013) pp. 816–819.
20. A. Nikfarjam and G. H. Gonzalez, Pattern mining for extraction of mentions of adverse drug reactions from user comments, in *AMIA Annual Symposium Proceedings*, 2011.
21. C. C. Yang, L. Jiang, H. Yang and X. Tang, Detecting signals of adverse drug reactions from health consumer contributed content in social media, in *Proceedings of ACM SIGKDD Workshop on Health Informatics (August 12, 2012)*, 2012.
22. J. Bian, U. Topaloglu and F. Yu, Towards large-scale twitter mining for drug-related adverse events, in *Proceedings of the 2012 international workshop on Smart health and wellbeing*, 2012.
23. A. Sarker and G. Gonzalez, *Journal of Biomedical Informatics* **53**, 196 (2015).
24. M. Yang, M. Kiang and W. Shang, *Journal of biomedical informatics* **54**, 230 (2015).
25. D. Li, N. Xia, S. Sohn, K. B. Cohen, C. G. Chute and H. Liu, Incorporating topic modeling features for clinic concept assertion classification, in *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, 2013.
26. H. Bisgin, Z. Liu, H. Fang, X. Xu and W. Tong, *BMC bioinformatics* **12**, p. S11 (2011).
27. A. K. McCallum, Mallet: A machine learning for language toolkit, http://mallet.cs.umass.edu, (2002).
28. D. Newman, A. Asuncion, P. Smyth and M. Welling, *The Journal of Machine Learning Research* **10**, 1801 (2009).
29. R. Ginn, P. Pimpalkhute, A. Nikfarjam, A. Patki, K. O'Connor, A. Sarker and G. Gonzalez, Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark, in *proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM). Reykjavik, Iceland*, 2014.
30. M. Paul and R. Girju, *Urbana* **51**, p. 61801 (2010).
31. I. Titov and R. McDonald, Modeling online reviews with multi-grain topic models, in *Proceedings of the 17th international conference on World Wide Web*, 2008.
32. R. Y. Lau, S. Liao, R. C. W. Kwok, K. Xu, Y. Xia and Y. Li, *ACM Transactions on Management Information Systems* **2**, 1 (2011).