SideEffectPTM: An Unsupervised Topic Model to Mine Adverse Drug Reactions from Health Forums

Sheng Wang, Henry Lin, Xuefeng Zhu, Jason Cho, Shantanu Dev,
Cassandra Spirit Jacobs, Ran Meng, Stephen Monette
Department of Computer Science
University of Illinois at
Urbana-Champaign
Urbana, IL 61801, USA
swang141@illinois.edu

ABSTRACT

Drug side effects is the fourth leading cause of death in the U.S. after heart failure, cancer and stroke. Clinical trials are not able to discover all drug side effects due to the limited patients and limited time. The post marketing drug safety mainly relies on spontaneous reports to The Food and Drug Administrations (FDA). However, the FDA's database is hard for patients to use because of the significant mismatch between patient terminology and standard medical vocabularies. Furthermore, it may take years for the FDA to decide to withdraw any dangerous drugs. In this project, we try to discover side effects by mining the growing online health forum data where many patients voluntarily report detailed cases about suspicious side effects.

1. INTRODUCTION

Automatic discovery of medical knowledge using data mining has huge benefit in improving population health and reducing healthcare cost. Applying data mining to discover adverse drug reaction (ADR) [4, 6, 8, 9, 16] is especially important because of the huge damage of ADRs to patients. An adverse drug reaction refers to the harm associated with the use of given medication at a normal dosage during normal use [11]. It is the fourth leading cause of death in the U.S., following heart disease, cancer and stroke [5]. A recent survey shows that ADRs account for 5% of all hospital admissions, occur in 10-20% of hospital inpatients and cause deaths in 0.1% of medical and 0.01% of surgical inpatients [1, 2, 14]. Approximate $136 billion has been spent on treating ADRs each year in U.S. [7, 13]. Besides these shocking statistical results, ADRs also affect patients' quality of life, cause patients to lose confidence in their physicians and increase costs of patient care [12]. ADRs would also mimic disease, resulting in unnecessary investigations and delay in treatment [12].

Although clinical trials are used to examine drug safety before drugs reach the market, they can hardly uncover all ADRs due to the limited patients and the limited time. The post marketing drug safety mainly relies on spontaneous reports from hospitals, pharmaceutical companies and patients to the Adverse Event Reporting System (AERS). However, it may take years for The Food and Drug Administrations (FDA) to withdraw or restrict dangerous drugs based on these reports. For example, trovafloxacin, which went on the market in Feb 1998 as a broad spectrum antibiotic, was not withdrawn until June 1999 due to the risk of hepatotoxicity.

The drug producers are thus seeking for a more timely approach to detecting the potential danger of their new drugs. Patients and doctors also need more informative and detailed cases about ADRs instead of a mere drug-symptom pair. In addition, the AERS is often difficult for patients to use. The significant mismatches between patient terminology and both the information source terminology and standard medical vocabularies often lead to confusion and misunderstanding [9, 17]. In this paper, we explore a promising alternative way of discovering ADRs by mining the growing online health forum data where many patients voluntarily report detailed cases about suspicious ADRs.

Recently, the dramatic growth of social networks has changed our lives in many ways, especially in how people interact with each other, how people obtain information and how they solicit opinions. In particular, as ad hoc social networks, online health forums have attracted more and more patients to describe the ADRs they experienced and seek for help [4, 10, 15, 16]. Online health forums contain the richest detailed text information about patients' ADRs as well as other patients' similar experience in comparison to drug review systems [9] and general social networks [6]. Patients with serious ADRs are more likely to post their conditions on an active health forum than on a noisy social media. The wide coverage of drugs and abundant patients' personal experience make online health forum a clear valuable source of mining previously unknown ADRs. In this paper, we study how to mine ADRs form online health forums. Adding to the current practice of AERS reporting, this new way of discovering ADRs can potentially allow us to discover ADRs much faster, and may also be advantageous in discovering long-term or rare ADRs and drug-drug interactions. In addition, we can link the mined ADRs to the related threads in health forums to show more detailed personal cases to users in order to compensate the brief information from AERS.

2. MAJOR FUNCTIONAL MODULES

We introduce three major novelty functional modules in this section.

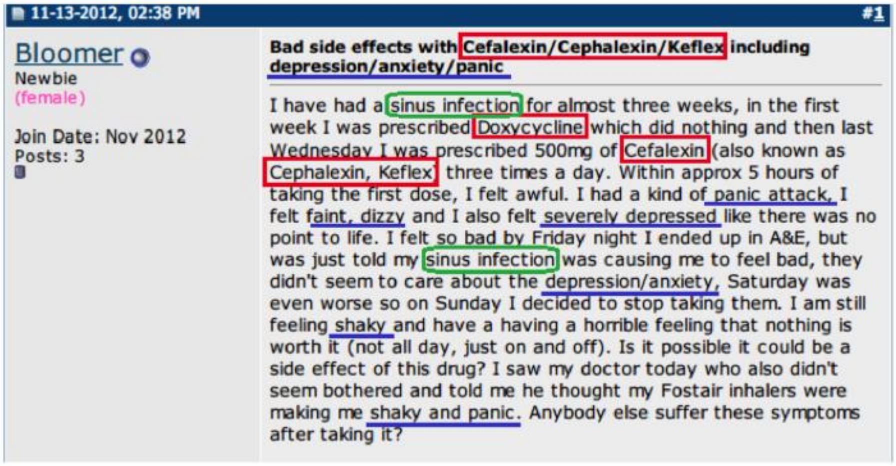
2.1 Data Preprocessing

We crawl data from a real-world online health forum HealthBoards (www.healthboards.com). HealthBoards is an online health forum that allows patients to discuss their conditions. It has been rated as one of the top 20 health information communities, with over 10 million monthly visitors, 850,000 registered members and over 4.5 million messages posted. We crawl a large text collection of 330,305 threads. We then extract 886 threads from a board called "Drug Interation/Side Effects". Existing NLP tools perform reasonable well when we extract drugs from the raw text. We use Metamap [3] to extract 287 drugs from all the 886 documents. Each thread has on average 336 words. As we have mentioned before, each thread may describe more than one drug. In our data, we find an average of 3.18 drugs per thread. We filter the stopwords in the text and use a large medical phrase dictionary to extract all the medical related phrases. We also remove the high-frequency and low-frequency phrases. In total, we build a dictionary with 2107 phrases.

We use the large corpus of 330,305 threads to build the phrases distribution of the background topic. We also use these 330,305 threads to build a unigram model as a prior for the treated symptom topic of each kind of drug. The treated symptom topic would be further updated by the training collection.

2.2 Mining ADRs

Mining ADRs from online health forums is challenging because of the complicated user-generated text. Patients may describe more than one drug and more than one symptom in one thread. The drug and the corresponding side effects are not restricted to be in the same thread. Not all the extracted symptoms are ADRs. Some of them could be the treated symptoms which patients want to treat by using that drug. Therefore, we need to separate these different symptoms in the mined results. Besides, we need to use an unsupervised approach since we don't have any human labeling data.



11-13-2012, 02:38 PM #1

Bloomer
Newbie
(female)

Join Date: Nov 2012
Posts: 3

Bad side effects with Cefalexin/Cephalexin/Keflex including depression/anxiety/panic

I have had a sinus infection for almost three weeks, in the first week I was prescribed Doxycycline which did nothing and then last Wednesday I was prescribed 500mg of Cefalexin (also known as Cephalexin, Keflex three times a day. Within approx 5 hours of taking the first dose, I felt awful. I had a kind of panic attack, I felt faint, dizzy and I also felt severely depressed like there was no point to life. I felt so bad by Friday night I ended up in A&E, but was just told my sinus infection was causing me to feel bad, they didn't seem to care about the depression/anxiety, Saturday was even worse so on Sunday I decided to stop taking them. I am still feeling shaky and have a having a horrible feeling that nothing is worth it (not all day, just on and off). Is it possible it could be a side effect of this drug? I saw my doctor today who also didn't seem bothered and told me he thought my Fostair inhalers were making me shaky and panic. Anybody else suffer these symptoms after taking it?

Figure 1 An example of a forum thread. The red rectangles refer to the drugs. The green ellipses refer to the treated symptoms. The blue lines refer to the side effect symptoms

To solve these challenges, we propose a novel unsupervised topic model which can explicitly build different topics for different kinds of drugs. Similar to "bag of words" assumption, we make a "bag of symptoms" assumption to model the text with more than one symptom and more than one drug. We incorporate prior to align the topics to treated symptoms and side effect symptoms. We finally use Maximum A Posterior for parameter estimation.

When user input a drug name, we will return the top ADRs mined from online health forum. These ADRs are ranked according to our unsupervised topic model. We show the results of several drugs of our system in table 1. From table 1, we see that we successfully mined many meaningful ADRs based on our model.

2.3 Bridging the language gap

There are language gaps between patient language and doctor language [17]. Patients and doctors tend to use different phrases describe similar symptoms based on their medical background. For example, patient may describe their symptoms as “difficulty breathing” while doctor may use more precise word “dyspnoea”.

To bridge the language gap, traditional approaches use a word level translation between patient words and doctor words. In comparison to these methods, we solve this problem in a topic perspective. We directly model both distribution on doctor words and distribution on patient words in our unsupervised topic model. In this way, we can show the sorted ADRs lists in both patient language and doctor language.

Table 2 is an example of the patient mode ADRs and doctor model ADRs for drug Zoloft. Here PLSA is the baseline method for patient mode. PLSA+SIM is the baseline method for doctor model. NPLSAWis our method for patient model. NPLSAC is our method for doctor model. We see that our result contain less noisy words and is more interpretable.

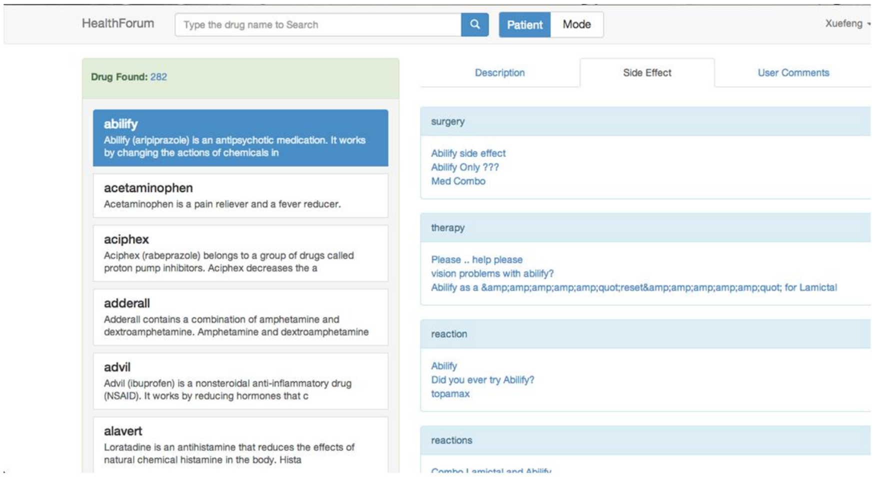
We allow users to switch between having the side effects change from side effect terms coined by doctors with side effects coined by patients.

2.4 Linking ADRs to forum posts

There are many existing sites that provide the information about ADRs to a given drug, including www.drugs.com, www.fda.gov, www.webmd.com. However, none of these sites will further link the ADRs to the forum posts that discuss them. A more detail posts describing the personal experience with that ADRs can help users better understand the various dangers. Therefore, we add this functional module which links the mined ADRs to the related forum posts.

We model this problem as an information retrieval task where the ADRs are the queries and the related posts are the documents. We design a specific retrieval function which can emphasize the ADRs in the forum posts. Moreover, since we build ADRs in both patient language and doctor language, it is desirable if we can retrieval the posts in both language. Retrieving in patient language is easy to achieve, since forum posts are written by patients. To retrieve in doctor language, we use a query expansion technique which expand the query in doctor language to patient language based on their semantic similarity.

Figure 2 is an example of our functional module that link the ADRs to the forum posts.



Drug Found: 282

abilify
Abilify (aripiprazole) is an antipsychotic medication. It works by changing the actions of chemicals in

acetaminophen
Acetaminophen is a pain reliever and a fever reducer.

aciphex
Aciphex (rabeprazole) belongs to a group of drugs called proton pump inhibitors. Aciphex decreases the a

adderall
Adderall contains a combination of amphetamine and dextroamphetamine. Amphetamine and dextroamphetamine

advil
Advil (ibuprofen) is a nonsteroidal anti-inflammatory drug (NSAID). It works by reducing hormones that c

alavert
Loratadine is an antihistamine that reduces the effects of natural chemical histamine in the body. Hista

surgery
Abilify side effect
Abilify Only ???
Med Combo

therapy
Please .. help please
vision problems with abilify?
Abilify as a &amp;amp;amp;amp;quot;reset&amp;amp;amp;amp;quot; for Lamictal

reaction
Abilify
Did you ever try Abilify?
topamax

reactions
Combo Lamictal and Abilify

Figure 2 An example of our functional module which link the ADRs of drug “zoloft” to the related forum posts.

3. OTHER COMPONENTS

Besides these three main modules, we also explore some other questions.

Table 1: Top-15 Phrases of Side Effect Symptom Topic

Drug(Freq)	Drug Use	Symptoms in Descending Order
zoloft(84)	antidepressant	**weight gain**, **weight**, depression, side effects, mgs, **gain weight**, anxiety **nausea**, **head**, brain, **pregnancy**, **pregnant**, **headaches**, depressed, **tired**
wellbutrin(48)	antidepressant	wellbutrin, Wellbutrin, **seizures**, depression, **seizure**, **sleep**, mgs,**weight loss**, period, **enery crashes**, pharmaceutical, high doses, dosages, crash, depressed
ativan(33)	anxiety disorders	ativan, **sleep**, seroquel, doc prescribed seroquel, **raising blood sugar levels**, anti-psychotic drug, **diabetic**, **constipation**, **diabetes**, 10mg, benzo, **addicted**, Ativan, plans, **vertigo**
topamax(20)	anticonvulsant	Topamax, **liver**, side effects, **migraines**, **headaches**, **weight**, topamax, pdoc, neurologist, supplement, **sleep**, **fatigue**, **seizures**, **liver problems**, **kidney stones**
zocor(3)	lipid-lowering agent	small pill, membranes, adverse reactions, **muscle cramps**, peripheral neuropathy, **memory problems** **memory loss**, heart palpitations, healthy diet, reactions, **vomiting**, tablets, **anemia**, **dizziness**, **dizzy**
ephedrine(2)	stimulant	**dizziness**, **stomach**, benadryl, **dizzy**, **tired**, **lethargic**, tapering, **tremors** **panic attack**, **head**, pshaw, advil cold, **stomach symptoms**, **cold sweats**, bpm

Table 2: Comparison of Drug Side Effect Topic

Zoloft				Ativan			
PLSA	PLSA+SIM	NPLSAW	NPLSAC	PLSA	PLSA+SIM	NPLSAW	NPLSAC
depression	depression	**weight**	weight decreased	**ativan**	anxiety	mood	sleep disorder
side effects	nausea	weight loss	memory impairment	anxiety	nervousness	drowsiness	anxiety
weight	insomnia	depression	depressed mood	**side effects**	nausea	sleep	nausea
anxiety	migraine	weight gain	disturbance in attention	**benzo**	headache	insomnia	confusional state
mgs	constipation	gain weight	vision blurred	sick	migraine	fatigue	memory impairment
head	headache	**damage**	weight increased	**Prozac**	insomnia	panic disorder	depressed mood
brain	cough	insomnia	menstruation irregular	addicted	cough	anxiety	multiple sclerosis relapse
tired	anaemia	panic disorder	activities impaired	**10mg**	depression	epilepsy	nervousness
dizziness	arthritis	fatigue	dry mouth	**drinks**	chest discomfort	depression	amnesia
depressed	dizziness	depressed	hot flush	addictive	arthritis	symptom	suicidal ideation
nausea	pain	**symptom**	stress	**therapy**	pain	**disease**	insomnia
weight gain	hypertension	high blood pressure	aggression	sleepy	fatigue	hypertension	malaise
wean	anxiety	dizziness	fall	vertigo	constipation	migraine	gait disturbance
headaches	pain in extremity	stress	condition aggravated	**antidepressant**	irritability	sleeping	extrapyramidal disorder

Anxiety and ADRs: What role will anxiety play when user read a list of potential ADRs? What kind of patient would be affected more by anxiety? Anxieties have been discovered to be appeared on search engine data. Is there also anxiety on health forum. Health forum is different from search engine that user could spread their feelings through discussion. Will anxiety spread through these discussions?

Measuring similarity between two ADRs: How to measure the similarity between two ADRs is a key factor in our model. Besides using existing APIs which provided similarity measurement between phrases, we explore other alternative ways. We use ADRs as input query to the existing search engines. We use the return snippets, links and documents as corpus to conduct semantic similarity comparison between these two queries.

Intelligent Robot: We build an intelligent robot which can automatically crawl new posts about the ADRs and drugs. If a question related to ADRs is posted in the forum, our robot will answer the question with some basic function information from our sites. We also insert a link in the posts to let users visit our sites for more information.

Demo System: We build both demo system in web app and android app.

6. CONCLUSIONS AND FUTURE WORK

We emphasize that not all of our features were implemented in the presented version of our website. For example, we obviously did not have enough users to create an online ecosystem. This, in turn, stopped us from being able to rank drug side effects.

However, our system has much room to grow. As my group discussed with Professor Zhai, we could list drug interactions with the given drug side effects. This would allow patients to understand whether a drug’s side effect is truly caused by a drug, or whether it’s caused by the interactions of two drugs he or she is consuming.

Furthermore, we could extend the personalization feature by ranking side effects based upon user data. For example, if a user securely posted his or her patient history online, we could create a system that correlates features of the patient histories with the side effects, effectivly predicting the likelihood a patient would suffer from a given side effect.

The last extension of our website is to allow patients to release his or her patient history publicly with ease.. One common problem in health informatics is the availability of data. Clinics that house medical records place certain restrictions on their data to protect the privacy of their patients, even though of them already agree to allow their data to be released. These inconveniences slow down the rate of research, which bars us from quickly making scientific breakthroughs.

On the other hand, if this system becomes online, allowing patients to quickly allow their medical records to be released publicly without many restrictions, we could make progress in the scientific community as a whole.

6. APPENDIX -- TEAM CONTRIBUTIONS

The team roles were as follows:

- **Sheng Yang** and **Jason Cho** helped extract the data from the forums. Sheng's research with Professor Zhai is based upon this project, and so he is the main spokesman of this project. Jason Cho is consulting for our group.
- **Henry Lin** helped manage each of the other developers in our group. He also built the webserver, managed the heroku server and mysql database, helped create the schema for the database, and inserted most of the data into the database. He also created the test framework for which we used to test POST and GET requests to our server.
- **Xuefeng Zhu** managed the interface of the website. Help design the architecture of the system, and brainstorm new ideas and functionalities. Mainly worked on the frontend, including designing the user interface and writing functional code. Collaborated with backend team for RESTful API design and helped debug server side code.
- **Ran Meng** constructed the automatic answer robot, which is used to crawl relevant posts and thread addresses, detect and retrieve drug side effects discussion threads, and to leave a message to attract people coming to our forum. The results are stored in a json file, which is then loaded into the mysql database.
- **Shantanu Dev** scraped drug information from other websites using Beautiful Soups. He also created the first parsers for Sheng's data. He helped parse the ranked drugs and side effects and helped populate the mysql database. In addition, he helped write the initial server, locally, so that it could post and request information from the mysql database, also running locally. He wrote a parser to fill in the various blanks that existed in the data, as much of the mined data did not have any relevance or was misspelled. In addition, he wrote a script to crawl through websites and extract the drug descriptions and prices.
- **Stephen Monette**, an offline student situated in Washington D.C., had the most experience in database work, out of all of us. He helped create the schema of the database, and also assisted in writing parsers to insert Sheng and Jason's data into the database. He also initiated work to convert the site to use a RESTful data model & API.
- **Cassandra Spirit Jacobs** wrote more test cases for our app, using Henry's framework. She wrote several scripts to access and display the data, mostly for testing purposes. These scripts queried the database using several different tests to return results indicating if the test passed, if there were minor issues or if the database itself was down or broken. Wrote tests to check that all data was returned, check for dummy data, verify empty indexes in both patient and doctor results, and check for data consistency.

7. REFERENCES

[1] D. W. Bates and et al. Incidence of adverse drug events and potential adverse drug events. JAMA: the journal of the American Medical Association, 274(1):29–34, 1995.

[2] D. W. Bates and et al. The costs of adverse drug events in hospitalized patients. JAMA: the journal of the American Medical Association, 277(4):307–311, 1997.

[3] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl 1):D267–D270, 2004.

[4] B. W. Chee, R. Berlin, and B. Schatz. Predicting adverse drug events from personal health messages. In AMIA, volume 2011, page 217. American Medical Informatics Association, 2011.

[5] K. M. Giacomini and et al. When good drugs go bad. Nature, 446(7139):975–977, 2007.

[6] 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, pages 117–125. Association for Computational Linguistics, 2010.

[7] R. Leone and et al. Drug-related deaths. Drug Safety, 31(8):703–713, 2008.

[8] M. Lindquist and et al. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the who international database. Drug Safety, 23(6):533–542, 2000.

[9] J. Liu, A. Li, and S. Seneff. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. In IMMM 2011, pages 91–96, 2011.

[10] X. Liu and H. Chen. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In Smart Health, pages 134–150. Springer, 2013.

[11] J. R. Nebeker, P. Barach, and M. H. Samore. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. Annals of internal medicine, 140(10):795–801, 2004.

[12] M. Pirmohamed, A. M. Breckenridge, N. R. Kitteringham, and B. K. Park. Fortnightly review: adverse drug reactions. BMJ: British Medical Journal, 316(7140):1295, 1998.

[13] C. S. van Der Hooft and et al. Adverse drug reaction-related hospitalisations. Drug Safety, 29(2):161–168, 2006.

[14] T.-Y. Wu and et al. Ten-year trends in hospital admissions for adverse drug reactions in england 1999–2009. Journal of the Royal Society of Medicine, 103(6):239–250, 2010.

[15] 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, pages 816–819. Springer, 2013.

[16] A. Yates, N. Goharian, and O. Frieder. Extracting adverse drug reactions from forum posts and linking them to drugs. In Proceedings of the 2013 ACM SIGIR Workshop on Health Search and Discovery, 2013.

[17] Q. Zeng, S. Kogan, N. Ash, R. Greenes, and A. Boxwala. Characteristics of consumer terminology for health information retrieval. Methods of information in medicine, 41(4):289–298, 2002.