IDENTIFYING SERENDIPITOUS DRUG USAGES FROM PATIENT-REPORTED MEDICATION OUTCOMES ON SOCIAL MEDIA

by

Boshu Ru

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2018

Approved by

_____
Dr. Lixia Yao

_____
Dr. Yaorong Ge

_____
Dr. Mirsad Hadzikadic

_____
Dr. Jing Yang

ABSTRACT

BOSHU RU. Identifying serendipitous drug usages from patient-reported medication outcomes on social media. (Under the direction of DR. LIXIA YAO.)

Drug repositioning has prominent advantages of lower safety risk and development cost than developing new drugs. It has attracted broad interests from the biomedical community. In the past decades, computational approaches have examined biological, chemical, literature, and electronic health record data for systematic drug repositioning. But due to the limitations of these data sources, neither of them alone appear sufficient for drug repositioning research. In recent years, more and more patients go to social media to report and discuss their medication outcomes. Of these reports, we noticed mentions of serendipitous drug usages, which we hypothesize to be new, independent data to study drug repositioning, in the sense of complementing other existing data sources to identifying and validating drug repositioning hypotheses.

In our first work, we examined medication outcome information available on four social media sites, namely WebMD, PatientsLikeMe, YouTube, and Twitter. We found the patient health forum of WebMD the best social media site for our research in terms of data availability and quality, but colloquial patient language is challenging for computers to process. In the second phase of dissertation, we explored state-of-the-art natural language processing (NLP) and machine learning methods to identify mentions of serendipitous drug usages in social media text. We curated a gold-standard dataset based on filtered drug reviews from WebMD. Among 15,714 sentences in total, our annotators manually identified 447 sentences mentioning novel desirable drug usages that were not listed as known drug indications by WebMD and thus were considered serendipitous drug

usages. We constructed features using NLP methods and medical knowledge. Then we built SVM, random forest, AdaBoost.M1, and deep learning models and evaluated their prediction power on serendipitous drug usages. Our best model (AdaBoost.M1) achieved an AUC score of 0.937 on the independent test dataset, with the precision equal to 0.811 and the recall equal to 0.476. Our models predicted several serendipitous drug usages, including metformin and bupropion for obesity, tramadol for depression and ondansetron for irritable bowel syndrome with diarrhea, which were also supported by evidences from scientific literature. These results demonstrated that patient-reported medication outcomes on social media are complementary to other data sources for drug repositioning. NLP and machine learning methods make this new data source feasible to use. In the end, we implemented NLP and machine learning methods explored in this dissertation to an open source software application for users without intensive NLP and machine learning skills to extract serendipitous drug usages mentioned in social media text.

ACKNOWLEDGMENTS

There are many people I would like to thank because I would not go this far without their helps.

First of all, I must express my gratitude and appreciation to my advisor, Dr. Lixia Yao, from whom I received consistent support, constructive criticism, and valuable instruction on my research and career development.

I would like to thank all other members in my dissertation committee – Dr. Yaorong Ge, Dr. Mirsad Hadzikadic, and Dr. Jing Yang. I have received valuable instruction from them during my PhD training. Without their kindness and help, I would not be able to complete my dissertation.

Furthermore, I thank Dr. Dingcheng Li, Dr. Huayu Li, Dr. Yong Ge, Charles Warner-Hillard, Kimberly Harris, and Madhuri Ratna Maddipatla, whom I had collaborated with in my research.

Also, special thanks are given to Ms. Sandra Krasuse and Ms. Kathleen Dunn for keeping the logistics and many other things peace of mind throughout my Ph.D. study, and to the department of software and information system for providing me learning resources and opportunities.

Last but not the least, I am very grateful to my family and friends. My parents and my significant other consistently support my study all these years. They are my very important reason for working hard. My friends Yueqi Hu, Chuqin Li, Dr. Tingting Li, Dr. Yuemeng Li, Junjie Shan, Jingyi Shi, Dr. Yue Wang, Dr. Jinyue Xia, Dr. Xiongwei Xie, and many others who helped me so much that words would fail to convey my gratitude.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Background

Drugs are chemical substances developed to treat or prevent diseases or improve the health status of human body. The development of a new drug includes discovery, design, clinical trials, and registration phases. It typically costs hundreds of millions of dollars and takes 10 to 17 years in total, with an average success rate of less than 10% [1]. Sometimes after a drug hits the market, it may be found useful to treat medical conditions other than what it is initially designed for. This strategy is known as drug repositioning or drug repurposing [2]. As opposed to the development of a new drug, repositioning a drug already in the market or in the late phases of development can save a considerable amount of time and financial resource, because the repositioned drug already passed several preclinical tests in animal models and safety tests on human volunteers in the Phase I clinical trials. Therefore, repositioning drugs are more available to patients of currently not properly treated diseases and more cost-effective to pharmaceutical companies [3]. A well-known example is sildenafil that was originally developed to treat angina. After the clinical trials on angina became futile, the clinical team found that some patients were reluctant to return the medicine because of the desirable side effect of erection [4]. This serendipitous finding inspired the team to explore the possibility of resurrecting the drug to treat erectile dysfunction and finally brought to the world the blue pill that are used by millions of men today [1].

Attracted by legendary stories like this, the biomedical community have examined various computational drug repositioning approaches in the past decades, using biomedical, literature, and electronic health record (EHR) data. But each of these data

sources has its own limitations, such as the limited capability to be translated to human patients, and thus has showed high false positive rates during predictions (See 1.2.1).

More recently, the fast expansion of social media generated a large amount of data. Many people write online to share their medication experience. As the data is generated from real patients, the translational hurdle from cell-line or animal model to human is bypassed. In this sense, social media data may be used to enrich or validate the drug repositioning signals generated from other existing data sources. For example, if a new drug usage is suggested by chemical or biological data, and we observe similar serendipitous drug usage from social media data, we would be much more confident about this repositioning opportunity and expect the chance of false positive discovery be much lower. However, how to use social media data for drug repositioning purpose has not been thoroughly investigated yet. Therefore, we proposed this dissertation study to design an information system that could automatically identify serendipitous drug usages, the important clues for drug repositioning, in social media data. The success of this system will be a valuable resource to the biomedical community and will contribute to the identification of new drug repositioning ideas.

1.2 Related work

This study complements existing computational methods for the drug repositioning and extends other medication outcome studies that utilized social media. This section gives a detailed discussion of the related work.

1.2.1 Computational drug repositioning methods

In the past decades, various computational methods have been developed to systematically generate more drug-repositioning hypotheses [2]. The basic idea is to mine

chemical, biological, or clinical data for drug similarity, disease comorbidity, or drug-disease associations that imply repositioning opportunities [2, 5]. For instance, Keiser *et al.* compared chemical structure similarities among 3,665 drugs and 1,400 protein targets to discover unanticipated drug-target associations and implicated the potential role of Fabahistin, an allergy drug, in treating Alzheimer's disease [6]. Sanseau *et al.* investigated data from genome-wide association studies to systematically identify alternative indications for existing drugs and suggested repositioning denosumab, which was approved to treat osteoporosis, for Crohn's disease [7]. Hu *et al.* created a drug-disease network by mining the gene-expression profiles in GEO datasbase and the Connectivity Map project [8]. By analyzing topological characteristics of this network, they inferred the effects of cancer and AIDS drugs for Huntington's disease. Wren *et al.* constructed a network of biomedical entities including genes, diseases/phenotypes, and chemical compounds from MEDLINE [9], and computationally identified novel relationships between those biomedical entities in scientific publications [10]. One such relationship they found and validated in the rodent model was between chlorpromazine and cardiac hypertrophy. Nevertheless, Gottlieb *et al.* designed an algorithm called PREDICT, to discover novel drug-disease associations from OMIM, DrugBank, DailyMed, and Drugs.com [11]. This algorithm predicted 27% of drug-disease associations in clinical trials registered with clinicaltrial.gov.

Although these computational methods have demonstrated their promise, they often face the issue of high false positive rates [2, 12]. One primary reason is sharing similar chemical structures or co-occurring in the same publication does not always imply medical relevance. Also, ignoring the context (e.g., whether the similarity or validation is

observed in experiments on molecular, cell line, or animal models) might impact their capability to be translated to human beings.

In addition to exploring novel ways of forming repositioning ideas, researchers recently began to validate drug-repositioning hypotheses in the EHR data. For example, Khatri *et al.* retrospectively analyzed the EHR data of 2,515 renal transplant patients at the University Hospitals Leuven to confirm the beneficial effects of atorvastatin on graft survival [13]. Xu *et al.* verified that metformin, a common drug for type 2 diabetes, is associated with improved cancer survival rate by analyzing the patients' EHR data from Vanderbilt University Medical Center and Mayo Clinic [14]. These proof-of-concept studies have also witnessed several limitations, due to the nature of EHR data: (1) EHR systems do not record the causal relationships between events (e.g., drugs and side effects) as they were mostly designed for clinical operation and patient management instead of research. Whether a statistical association is causal needs to be verified through temporal analysis with a lot of assumptions. Therefore, the models become disease and/or drug specific and remain difficult to be generalized and automated in a large scale. (2) A significant amount of valuable information, such as the description of medication outcomes, is stored in clinicians' notes in the free-text format [3]. Mining these notes requires advanced natural language processing techniques and presents patient privacy issues. (3) In the US, data from a single provider's EHR system only provide an incomplete piece of patient care [14]. Integrating EHR data from multiple providers may be a solution, but currently encounters legal and technical challenges, as discussed in depth by Jensen *et al* [15].

Due to these limitations, neither EHR, nor any of scientific literature, biological, and chemical data alone appears sufficient for systematically generating ideas for drug repositioning research. We need to identify additional data sources that contain patient medication history and outcomes, as well as develop advanced data integration methods to identify synergistic signals from multiple sources.

1.2.2 Medication outcome studies using social media data

In the last decade or so, social media data has increased exponentially in the volume. People today not only post their travel pictures but also share and discuss their experience with diseases and drugs on social media websites, such as WebMD, PatientsLikeMe, Twitter, and YouTube [16]. Such data directly describes drug-disease associations in real human patients and bypasses the translational hurdle from cell-line or animal model to human, thus they have led to increased research interests. For example, Yang *et al.* mined drug and adverse-reaction associations in the drug-related discussions on the MedHelp forum using the ADR lexicon generated from the Consumer Health Vocabulary (CHV) [17] and various metrics for evaluating associations [18]. They found that two ADR signal measures, namely *leverage* [19] and *PRR*, achieved better accuracy than the others when dealing with that their social media dataset. Yates *et al.* generated an ADR synonym set specifically for breast cancer patients from the United Medical Language System (UMLS), and used it to extract ADRs in the breast cancer drug reviews that they crawled from Askpatient.com, Drugs.com, and Drugratingz.com [20]. Instead of collecting available social media data, Knezevic *et al.* created a Facebook group for people to report their ADR experiences [21]. The group quickly attracted 973 Facebook users in the first 7 months. Moreover, 2% of the users reported ADRs, which is much

higher than the reporting ratio of several spontaneous reporting systems, demonstrating that social media is a highly sensitive instrument for ADR reporting. Powell *et al.* investigated the MedDRA Preferred Terms that appeared on Twitter and Facebook and found 26% of the posts contained useful information for post-marketing drug safety surveillance [22]. However, these studies relied on one single platform of social media or focused narrowly on specific ADR signals, leaving the potential use of social media for studying other aspects of medication outcomes, such as serendipitous drug usages unexamined.

Based on the recent research in computational repositioning methods and the efforts to use social media for medication studies, we believe that: (1) Additional data source, such as patient-reported medication history and outcomes can be helpful for generating and validating drug repositioning ideas. (2) The potential and possible solution of using social media for drug repositioning purposes needs to be further investigated.

1.3 Overview of this dissertation

This dissertation is organized in four parts. In Chapter 2, we surveyed four social media sites to identify best data source and challenges of mining medication outcome information [16]. In Chapter 3, we curated a gold-standard dataset based on filtered drug reviews from WebMD and built a natural language processing and machine learning pipeline to identify serendipitous drug usages in patient forum data. In Chapter 4, we applied cutting-edge word embedding and deep learning methods and discussed their performance in the context of this dissertation. In Chapter 5, we documented how we designed and implemented an open source software application based on natural language processing and machine learning methods explored in this dissertation work.

CHAPTER 2: A CONTENT ANALYSIS OF PATIENT-REPORTED MEDICATION OUTCOMES ON SOCIAL MEDIA

2.1 Background

According to a survey in 2013 [23], 30% of adults were willing to share their health information on social media. If we also take those early adopters more than a decade ago into account, social media should have accumulated huge amounts of health data. But how big are the data? What is the quality of those data? Are there any differences among data from different social media sites?

In this section, we compared and evaluated four major representative social media platforms, in terms of data coverage and quality. We examined what kinds of medication outcomes were discussed and investigated the characteristics of the informal written languages used online. Such work is necessary for us to have thorough understanding of patient-reported medication outcomes on social media, before we developed a computational system later to identify serendipitous drug usages from this new data source.

2.2 Methodology

2.2.1 Data collection

We selected four representative chronic diseases, namely asthma, rheumatoid arthritis, type 2 diabetes and cystic fibrosis. The former three are common diseases, which are expected to have a lot of patient reviews on social media; whereas, cystic fibrosis is a rare disease and is expected to have much less data available. For each of these diseases, we picked two to three commonly prescribed drugs to be studied. In total we got 11 disease-drug pairs (See Table 1). We did not pick severe disabling diseases or acute conditions,

Table 1: List of diseases and drugs

Disease-Drug Pair		Alternative Drug Names
Asthma	Albuterol	Ventolin, Salbutamol
	Ipratropium	Atrovent, Apovent, Ipraxa, Aerovent, Rinatec
	Prednisone	Deltasone, Prednicot, Rayos, Sterapred
Cystic Fibrosis	Azithromycin	Zithromax, Sumamed, Zmax, Azaste
	Ivacaftor	Kalydeco
Rheumatoid Arthritis	Meloxicam	Mobic
	Prednisone	Deltasone, Prednicot, Rayos, Sterapred
	Sulfasalazine	Azulfidine, Salazopyrin
Type 2 Diabetes	Bromocriptine	Parlodel, Cycloset, Bagren, Pravidel
	Insulin	Levemir, NovoLog, Lantus, Afrezza, Apidra,HumaLog, Humulin, Novolin, KwikPen
	Metformin	Fortamet, Glucophage, Glumetza

because patients with those diseases probably are not able to write online or lack of long and sustainable interests to write online.

The social media sites we surveyed include WebMD, PatientsLikeMe, YouTube and Twitter. Figure 1 demonstrates what a typical user review is like on each site. The former two sites are specialized in exchanging medical and health information, while the latter are two of the largest social network sites in the US. For WebMD and PatientsLikeMe, we collected the drug reviews from the drug and treatment category pages (URLs: http://www.webmd.com/drugs/index-drugs.aspx and https://www.patientslikeme.com/treatments). For YouTube and Twitter, we searched disease and drug names using their search APIs, and then parsed the retrieved comments using the JavaScript scraper written by us. To ensure the completeness of the results,

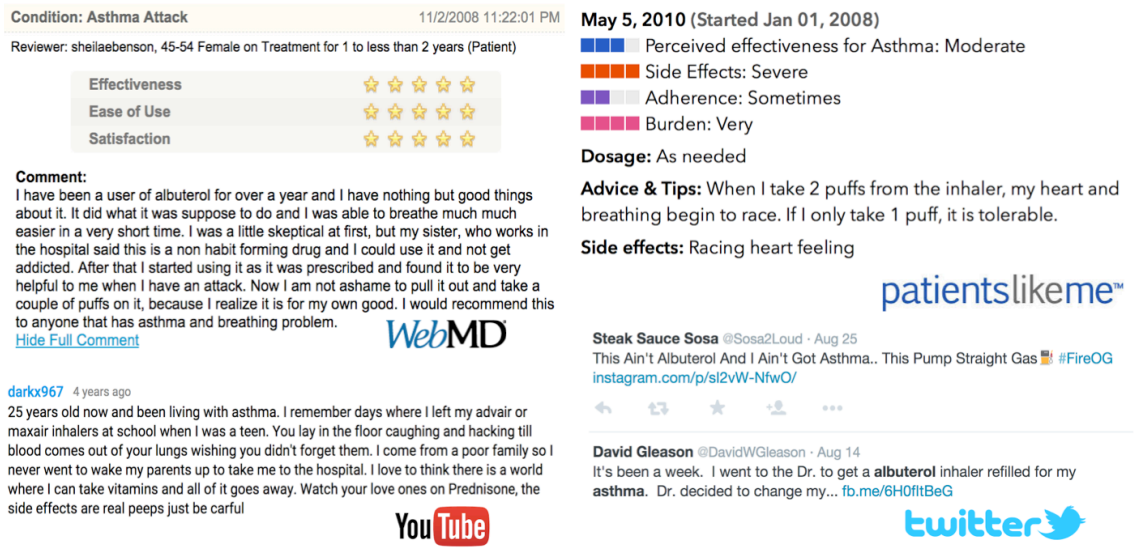

Figure 1: Examples of the drug reviews posted on WebMD, PatientsLikeMe, YouTube, and Twitter

disease and drug synonyms were also used to form the search queries. The alternative drug names were listed in the Table 1. Alternative disease names were found in UMLS. In this way we collected all the publically visible data published by October 1, 2014 on all four sites. Facebook was included in our initial evaluation but got dropped out of our final analysis, because most Facebook users discuss about personal medication experiences in private group setting.

2.2.2 Data preprocessing and indexing

The data we collected contained structured data and free-text comments. Structured data can be the user ratings on WebMD and PatientsLikeMe. WebMD uses a standard five-point rating system for users to rate on effectiveness, ease of use, and overall satisfaction. PatientsLikeMe asks users to choose one from several preset options - such as *Major*, *Moderate*, and *Slight* - to describe the drug's effectiveness, side effects, adherence, and burden. To make the data from these two sites comparable, we

proportionally converted the rating values from PatientsLikeMe into the five-point system. For all the free-text reviews, we used Apache Lucene, a free open source information retrieval software to index them (http://lucene.apache.org/core/4_10_0/) based on the following medication outcome lexicon.

2.2.3 Medication outcome lexicon and categorization

We built a medication outcome lexicon to identify the related terms in social media. Our lexicon consists of four vocabularies representing four major medication-outcome categories: effectiveness, side effects, adherence, and cost. The vocabularies for effectiveness, adherence, and cost were manually collected from the WebMD corpus, as there are no existing terminologies. The vocabulary for side effects was extracted from the Consumer Health Vocabulary, but we used UMLS to limit the semantic types of those terms to *Disease or Syndrome, Finding, Sign or Symptom, Neoplastic Process, Injury or Poisoning,* and *Mental or Behavioral Dysfunction*. For terms in each vocabulary, we searched them in the indexed documents. If a match was found, the review post was tagged to the related medication outcome category. It is not uncommon for a review post to be indexed by multiple terms and assigned to multiple medication outcome categories.

2.2.4 Sentiment analysis

In order to understand how social media users felt about their medication experiences, we used Deeply Moving, a free tool developed by the Stanford NLP group [24] for sentiment analysis. Deeply Moving parsed each sentence in the input document into a tree structure, with each leaf node representing a word used in the original text. Then it used a pre-trained Recursive Neural Tensor Network model on a corpus of movie reviews, to annotate the input sentence with one of five sentiment tags: *Very Negative,*

Negative, Neutral, Positive and *Very Positive*. We summarized the distribution of sentiment tags in different medication outcome categories for all four social media sites.

2.3 Results

2.3.1 Summary of data

We collected 2,567 reviews from WebMD, 796 reviews from PatientsLikeMe, 42,544 comments from YouTube, and 39,127 posts from Twitter. The significantly higher numbers for YouTube and Twitter may be due to the fact that they target much bigger user populations of broader interests. Another possible reason is that unlike WebMD and PatientsLikeMe, YouTube and Twitter do not compile and edit the published contents unless they are against the law or their company policies. Table 2

Table 2: Summary of drug reviews on social media sites

Diseases and Drugs \ Social Media and Metrics		WebMD				PatientsLikeMe					YouTube	Twitter
		Effectiveness	Ease of Use	Satisfaction	No. of Reviews	Perceived Effectiveness	Side effects	Adherence*	Financial Burden*	No. of Reviews	No. of Posts	No. of Posts
Asthma	Albuterol	3.87	4.39	3.55	112	4.06	2.08	4.01	4.29	137	2859	11381
	Ipratropium	4.17	4.17	3.75	12	4.13	1.33	4	3.95	8	140	271
	Prednisone	4.00	3.92	3.32	367	4.19	2.89	4.68	4.17	48	8569	971
Cystic Fibrosis	Azithromycin	-	-	-	-	2.91	1.73	4.55	4.65	11	31	90
	Ivacaftor	-	-	-	-	-	-	-	-	-	818	5060
Rheumatoid Arthritis	Meloxicam	3.39	4.18	3.11	202	3.20	1.89	4.47	4.6	15	558	67
	Prednisone	4.11	4.31	3.61	229	4.03	3.20	4.68	4.17	63	10064	670
	Sulfasalazine	3.19	3.28	3.17	65	2.44	3.03	4.45	3.98	77	581	53
Type 2 Diabetes	Bromocriptine	2.23	3.08	2.15	13	-	-	-	-	-	29	214
	Insulin	3.50	4.22	3.35	265	4.21	1.60	4.73	4.2	106	11401	16308
	Metformin	3.29	3.91	2.93	1302	3.69	2.30	4.41	4.17	331	7504	4042
Weighted Average:		3.52	4.00	3.14	-	3.74	2.32	4.42	4.18	-	-	-

* The average drug adherence and burden (cost) ratings on PatientsLikeMe may not be disease specific. In case that a drug has multiple indications, the adherence and burden ratings are consolidated across all indications.

summarized the counts of user entries we collected on each site for each disease-drug pair. It is found with no surprise that the prevalence of a disease impacts its popularity on social media. For instance, cystic fibrosis is a rare disease that affects approximately 30,000 people in the United States [25]. Consequently, we found zero reviews about its treatment on WebMD (It is also possible that WebMD does not include azithromycin and ivacaftor as treatment for systic fibrosis), 11 reviews for azithromycin on PatientsLikeMe and limited posts on YouTube and Twitter. On the contrary, thousands of posts talked about type 2 diabetes, which affects more than 29 million people in the United States alone [26]. Table 2 also showed the average patient ratings on WebMD and PatientsLikeMe. The Pearson's correlation coefficient between two sites is 0.728 for the effectiveness rating and 0.759 for the adherence (approximately equivalent to ease of use) rating, which demonstrates that the ratings on these sites are quite consistent. In addition to this, 79.4% of reviews on WedMD come with free-text comments. The number for PatientsLikeMe, however, is only 2.8%. The design of their drug review forms might account for this difference. WebMD encourages users to write their medication experiences in their own language. PatientsLikeMe, on the other hand, promotes users to fill out a standardized questionnaire with many multiple-choice questions; users might skip the optional free-text field at the very end of the questionnaire.

2.3.2 Data quality by sites

To evaluate what patients really wrote about, we manually reviewed all the reviews on WebMD and PatientsLikeMe, and 500 randomly selected posts from both YouTube and Twitter. We found that almost all the comments on WebMD and PatientsLikeMe described patients' experiences with drugs. By contrast, only 9.4% of the YouTube

comments discussed drug-related personal experiences. About 23.2% posts were about disease or drug- related knowledge or commercials. The rest 67.4% of the posts were simply spams or discussions on completely irrelevant topics. On Twitter, only 1.6% of the tweets wrote about drug-related personal experiences, 93% of the sampled tweets were mentioning of disease education articles, drug related news and commercials, research publication announcement, and patient recruitment notices. The rest 5.4% tweets were talking about completely irrelevant topics. Overall speaking, WebMD has the patient reviews of the highest quality among four, followed by PatientsLikeMe. The social media giants, particularly YouTube and Twitter, are dubiously mature sources for studying medication outcomes unless the precision of information retrieval could be significantly improved. In case that large sample size is needed, WebMD, PatientsLikeMe and other medicine-focused social media sites, can be combined.

2.3.3 Types of medication outcomes discussed on social media

We then looked at what specific topics social media users talked about and with what attitudes using Lucene and Deeply Moving (See 2.2). These analyses were conducted on sentence level because patients often address more than one medication outcome in a review. Figure 2 summarized our findings. Side effects were the most frequently mentioned outcome on all four sites, followed by the effectiveness and adherence. Patients seem not very sensitive about the cost, presumably because most people have pharmacy coverage in their health plans. But it is hard to infer if patients care more about drugs' side effects than effectiveness based on the sentence counts because the numbers could be complicated by the psychological effect of *negativity bias* [27], a notion that, even when of equal intensity, negative things have stronger impact on

a person's impression and evaluation than positive things. Table 3 gives a few examples of user comments on drug effectiveness, side effects, adherence and cost. Figure 2 also illustrated the sentiment score that Deeply Moving [24] assigned to all the sentences in user reviews. The color overlay inside each bar tells the relative ratios of five sentiments, from very negative, negative, neutral to positive and very positive, in each medication outcome category for all four social media sites. The negative sentiment dominated patients' discussion on effectiveness, side effects, adherence and cost across all four sites. This is, again, common to user reviews in many other fields, due to negativity bias [28]. Second to that are neutral and positive sentiments. Extreme sentiments, either very positive or very negative, were rare. However, the weighted average of numerical ratings for all medication outcome types but side effects (Bottom row of Table 2) were higher

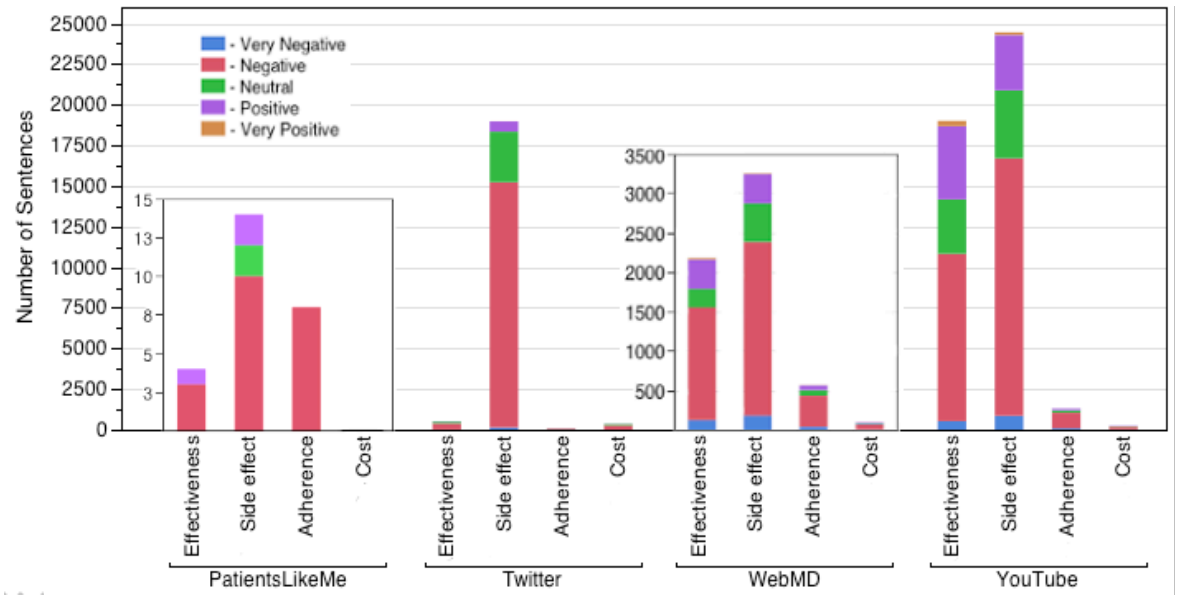

Figure 2: Summary of the sentiments associated with medication outcome types

(The numbers for PatientsLikeMe and WebMD are very small, so we adjust the scale marks for these two sites for better visibility.)

than 3.0 (neural). Possible explanations are either Deeply Moving was inaccurate for drug reviews, or patients used negative language and relatively large space to comment on side effects. But when they did the rating, they accepted the common fact that all drugs have side effects.

Table 3: Examples of medication outcome contents on social media sites

Content Type	Example	Source
Effectiveness	day 001 - ***not so bad so far, but really only 18 hours in***. Asthma acting up a little. Taking Advair. Need to refill Albuterol script.	Twitter
	Only been on it for 8 days but I noticed ***relief from pain*** right away. However, yesterday I was more short of breath than usual and my ***blood pressure was high***, tightness in my chest so I will stop taking it to see if that stops.	WebMD
Side effects	make sure it is plain claritin. not claritin D or anything else. ***Plain claritin has loratadine***. that is fine but anything else can kill it. ***the regular claritin has no side effects.***	YouTube
	Bubble gut to me is extreme gas... ***Very extreme gas, be careful for the first few weeks of use***. Don't put yourself in situations like elevators or long car rides before your body adjust to the med.	
Adherence	It is so very ***easy to use*** and the needle is tiny and does not hurt at all.	WebMD
	They divide pills down to ***easier to take*** portions or fine tune the dose to better fit the patient. Have you ever tried to use a breakfast drink with the pills? Sometimes it helps pills slide down MOBETTA!	YouTube
Cost	Private ***insurance pays*** for young Shan 's `miracle drug'; Kelsa can't ***afford*** it.	Twitter
	Being so close to Mexico allowed me to get my inhalers ***cheaper*** and quicker.	WebMD
Disease comorbidity & repositioning	Doctor prescribed this after I stopped taking Plaquenel due to stomach upset. In addition to RA I have a history of IBS, sensitive stomach and I have tolerated this medication well. ***It has greatly improved my IBS while moderately improving my RA pain***. Ony side effect is feeling full, thirsty and ocassional gut pain.	WebMD
	Personal side effect: Very sensitive to sun. ***This clears up my bronchial spasms so quickly and as a bonus, it clears up my eczema!*** I have asked my doctor to prescribe it regularly for my skin condition and he says there are too many side effects. too bad. It is a wonder drug. I was able to breathe very well.	

During our manual review, we noticed some interesting cases where the patients reported that some drugs unexpectedly helped with their comorbid conditions (See Table 3). For example, a couple of sulfasalazine users reported that their irritable bowel syndrome (IBS) symptoms were alleviated when taking this medicine for treating rheumatoid arthritis. This is possible because recent research and clinical trials found that sulfasalazine was able to relieve the diarrhea and abnormal pain that IBS patients often suffer from, act by stimulating CD73-dependent adenosine production [25]. We also found asthma patients on prednisone reported that the drug improved their eczema condition. Further literature search led us to find that prednisone has been linked to reducing the flare in atopic, seborrheic, and urticarial dermatitis, although only a few clinical studies have formally evaluated these off-label indications [29]. Such examples illustrated the potential value of social media data for studying the biological mechanisms of disease comorbidity and drug action, and repositioning existing drugs for new indications.

2.3.4 Challenges of analyzing the human language on social media

We also noticed that some usages of the written language on social media might be challenging for computers to process (See Table 4). First of all, informal writing conventions, typos and improper punctuation are widespread on social media websites. Special lexicon is needed to automatically recognize and correct those usages in social media data. Secondly, emoticons (i.e., ":o(" and "=0)"), exclamation marks, and uppercases give strong hints of the feelings, attitudes and opinions of the users. However, most text mining tools today do not capture those emoticons, extra exclamations marks, or

uppercases [30]. Thirdly, sarcasm detection is a particularly difficult data mining task and one possible solution is suggested by Tsur *et al.* for online product reviews [31]. Nevertheless, in many cases people did not provide direct opinions about a drug but instead compared it to other drugs. For example, a type 2 diabetes patient described his experience with metformin by comparing it to the glumetza treatment (See Table 4, the 1st example). In another case, corticosteroid was used in a YouTube comment to refer prednisone (See Table 4, the 7th example). The former concept is a drug class, of which the latter is a specific drug. While such comparisons are extremely useful information to

Table 4: Examples of the reviews challenging for processing

Challenge Type	**Examples**	**Source**
Comparative sentiment	Hugely intolerable diarrhea on a daily and nightly basis. ***The glumetza delivery solution is a bit better***	WebMD
	That being said, ***the only advantage Symbicort has over Advair is fometrol is faster acting.*** The LABA in Advair is matched in potency for the dosage in fometerol.	YouTube
Sarcasm	Quit taking it last week and I feel great. ***So I guess it worked by showing me how much worse I could feel.*** I'm still stiff and sore, but at least I don't feel like c***.	WebMD
	2 sick monkeys... Asthma and colds don't go well together. Albuterol does ***provide some comedy relief though #hypersilliness***. Prayers welcomed	Twitter
Informal language usage	***im*** taking 75 mg of prednisone high dose b/c of my kidneys and ***im*** concerned because ***im*** experiencing a lot of hair loss is this temporary or permanent??? My doctor will put me on a lower doses soon will that help with hair loss??	YouTube
	NPR coverage of Kalydeco from this am !	Twitter
Pronoun and semantic referencing	***Corticosteroids*** (a class of chemicals including Prednisone) causes too many collateral damage. The higher the dosage the more resistant your body becomes with insulin. Man I hate ***this medication*** (Prednisone), but is life saving to a certain extent.	YouTube
Emoticon	Muscle cramps, massive gastrointestinal upset. Who ever said this was the best drug for me to be taking was mad! The cure is worse than the illness! Apparently it will get easier the longer I take it. I've seen improvments but still can't leave the house at times ***:o(***	WebMD
	... I haven't had weight gain..(pls, I hope that doesn't change ***=0)***...lost 17 pounds in 2 months..however, began high protein, lowfat eating plan at same	

comparative effectiveness research, identifying the pronoun reference and semantic reference within a sentence or cross multiple sentences is still not properly solved. The ontology based reasoning will be needed for the machine to extract the valuable information from the patients' reviews correctly.

2.4 Discussion

In this work, we surveyed four major social media sites, namely WebMD, PatientsLikeMe, YouTube and Twitter to better understand if patients reported and discussed their personal medication experiences on social media and what the contents are like. By comparing the results for four carefully selected chronic conditions and 11 drugs, we found that in addition to consistent ratings, patients did share their feedback on effectiveness, side effects, adherence, and cost of drugs in a responsible way. YouTube and Twitters retrieved much more data. But specialized medicine-focused websites such as WebMD and PatientsLikeMe maintained the higher data quality.

Patients talked mostly about side effects, followed by effectiveness and adherence. They were not very sensitive to cost. In spite of the negative tones patients used on side effects, patients gave neural to positive ratings for effectiveness, adherence and cost for all 11 drugs. In addition, some patients even reported unexpected desirable indications, or serendipitous usage, which could be clinical evidences to study the mechanisms of drug actions and to identify novel opportunities for drug repositioning.

However, our findings need to be considered with the following factors taken into account: (1) Our lexicon for effectiveness, adherence and cost was created from all the patients' reviews we found from WebMD. It could be incomplete and lowered the recall of our information retrieval from the other three sites. (2) The sentiment analysis tool,

Deeply Moving, was trained on movie reviews. Considering the entertainment nature of movies and life-saving nature of drugs, our sentiment analysis results could be inaccurate and are worth further investigation. (3) As a preliminary study, we only surveyed 11 disease-drug pairs on four publicly accessible sites. The situations for acute conditions, severe disabling diseases, and private discussion sites might be different from what we observed in our results.

Despite of these limitations, this study suggests that social media, particularly the medicine-focused social media sites, is a promising data source. It is complementary to spontaneous reporting systems and EHR systems for understanding patient-reported medication outcomes. The serendipitous drug usages mentioned by patients are import clue for forming and validating drug repositioning hypotheses. If a drug and its serendipitous usage were observed to occur together from not only social media, but also EHR systems and spontaneous reporting systems, we should expect the chance of false positive discovery be much lower than the cases where the co-occurrences were found from only one or two data sources. To fully unlock the value of social media data for medication outcomes research, we build a natural language processing and machine learning pipeline in next chapter to mine social media for serendipitous drug usages.

CHAPTER 3: USING MACHINE LEARNING METHODS TO IDENTIFY SERENDIPITOUS DRUG USAGES IN PATIENT FORUM DATA

3.1 Background

In the second work, we build a computational pipeline based on machine learning methods to capture the serendipitous drug usages on the patient forum published by WebMD, which was reported to have high quality patient-reported medication outcome data. We expect this an extremely difficult machine learning task because: (1) User comments on patient forum are unstructured and informal human language prevalent with typographic errors and chat slangs. It is unclear how to construct meaningful features with prediction power; (2) the mentioning of serendipitous drug usages by nature is very rare. Based on our experience with the drug reviews on WebMD, the chance of finding a serendipitous drug usage in user posts is less than 3% (See 3.2). Therefore, we caution the audience that our objective in this work is not to build a perfect pipeline or a high-performance classifier, but to perform a feasibility check and identify major technical hurdles in the entire workflow.

3.2 Methodology

To identify serendipitous drug usages in patient forum data, we built the entire computational pipeline, which includes data collection, data filtering, human annotation, feature construction and selection, data preprocessing, machine learning model training and evaluation, as illustrated in Figure 3. Each module was built using standard tools and methods and was further described below.

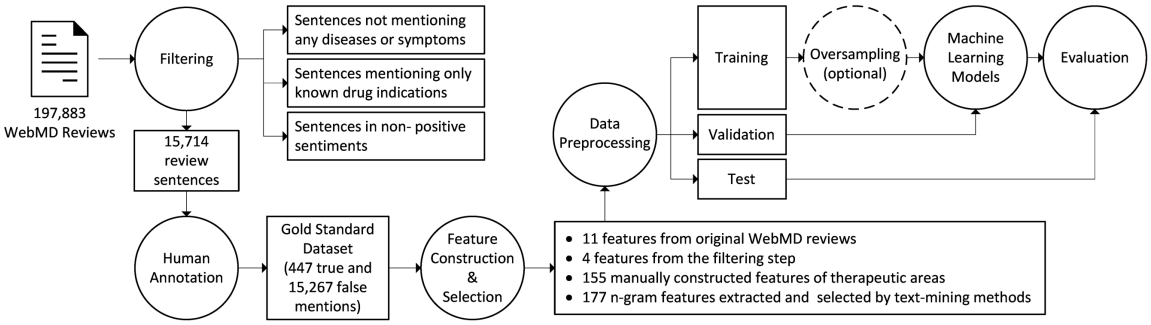

Figure 3: A workflow to identify serendipitous drug usages in patient forum data

3.2.1 Data collection

We started by collecting drug reviews posted by anonymous users on the patient forum hosted by WebMD. WebMD is a reputable health care website that exchanges disease and treatment information among patients and healthcare providers. In its patient forum, after filling the basic demographic information including gender and age group, users are allowed to rate drugs in terms of effectiveness, ease of use, overall satisfaction, and post additional comments about their medication experience (See Figure 4). We chose it based on two considerations: (1) With over 13 years' history of operation and on average over 150 million unique visits per month, WebMD contains a large volume of drug reviews that is highly desirable for conducting systematic studies. (2) The quality of drug reviews was reported to be superior to many other social media platforms in the previous study [16]. Spam reviews, commercial advertisements, or information irrelevant to drugs or diseases are rare, probably thanks to their forum modulators. We downloaded a total number of 197,883 user reviews on 5,351 drugs by the date of March 29, 2015. Then, we used Stanford CoreNLP [32] to break down each free-text comment into

sentences, which is the standard unit for natural language processing and text mining analysis.

3.2.2 Gold standard dataset for serendipitous drug usages

In machine learning and statistics, gold standard, or accurately classified ground truth data is highly desirable, but always difficult to obtain for supervised learning tasks. For identifying serendipitous drug usages, it would be ideal if a database of drug usages approved globally or customarily used off-label were readily available as the benchmark for known drug usages. The professional team at WebMD has published monographs to introduce each drug, including information on drug use, side effects, interactions, overdose, etc. We thus used such data as the benchmark for known drug usages in this work. We assume a drug use is serendipitous if the user mentioned improvement of his or her condition or symptom that was not listed in the drug's known indications according to WebMD (See the examples in Figure 4). Otherwise, we set the mentioned drug use to be non-serendipitous. Below we explain in more details how we applied this principal to

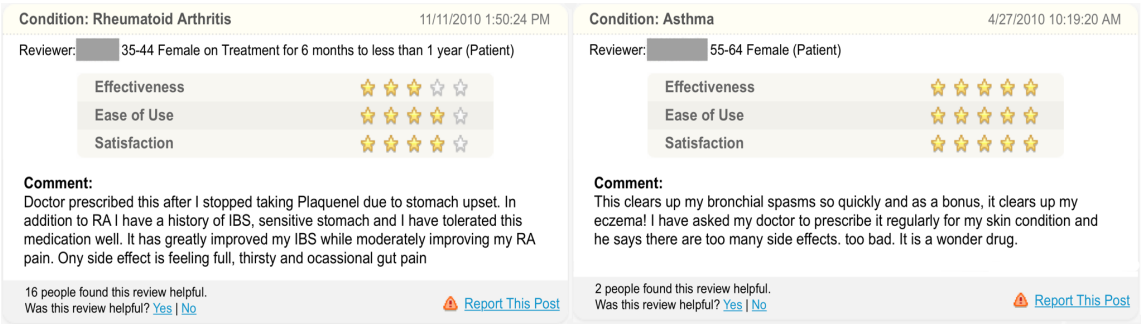

Figure 4: Examples of serendipitous drug usage mention on WebMD

In the example on the left, a patient reported that his irritable bowel syndrome (IBS) symptoms were alleviated when taking sulfasalazine to treat rheumatoid arthritis. In the example on the right, an asthma patient taking prednisone reported the improvement of her eczema.

semi-automatically prepare our gold standard dataset for serendipitous drug usages.

3.2.3 Data filtering

Three filters were designed to reduce the number of drug review sentences to a number more manageable for human annotation. Firstly, we identified and removed review sentences that did not mention any disease or symptom at all, because these sentences have no chance to be related to serendipitous drug usages. To do this, we selected the UMLS concepts in English and with the semantic types equal to *Disease or Syndrome*, *Finding*, *Injury or Poisoning*, *Mental or Behavioral Dysfunction*, *Neoplastic Process*, or *Sign or Symptom* and used them to approximate medical concepts that could be related to serendipitous drug usages. We then used MetaMap [33] to identify these medical concepts in each review sentence. Next, for sentences that did mention any of those concepts, we used SNOMED CT [34] to determine whether the mentioned concept is semantically identical or similar to the drug's known indications listed on WebMD. Mathematically SNOMED CT is a directed acrylic graph model for medical terminologies. Medical concepts are connected by defined relationships, such as *is-a*, *associated with*, and *due to*. The semantic similarity between two concepts was usually measured by the length of the shortest path between them in the graph [35, 36]. If the medical concept mentioned in a review sentence was more than three steps away from the known indications of the drug, we assumed the mentioned medical concept was more likely to be an unanticipated outcome for the drug and kept the sentence in the dataset for the third filter. Otherwise, we excluded the sentence from further evaluation, as it was more likely to be related to the drug's known usage rather than serendipitous usage we were looking for. In the third step, we used the sentiment analysis tool, Deeply Moving

[24] offered by the Stanford Natural Language Processing Group to assess the sentiment of each sentence where unanticipated medical concept occurred. We filtered out all sentences with *Very Negative, Negative,* or *Neutral* sentiment and only kept those with *Positive* or *Very Positive* sentiments because serendipitous drug usages are unexpected but desirable outcomes to patients. Negative sentiment is more likely to be associated with undesirable side effects or potential drug safety concerns. After these three filtering steps, 15,714 drug review sentences remained for human annotation.

3.2.4 Human annotation

One public health professional and one health informatics professional with master degrees, independently reviewed the 15,714 sentences and annotated whether each sentence was a true mention of serendipitous drug usage based on the benchmark dataset of known drug usages defined by WebMD. That is, they labeled a drug use to be serendipitous if the user mentioned an improved condition or symptom that was not listed in the drug's known indications according to WebMD. Otherwise, they assigned the mentioned drug use to be non-serendipitous. In case that the annotators did not agree with each other, they discussed and assigned a final label together. Six months later, the two professionals reviewed their annotation again to avoid possible human errors. In total, 447 or 2.8% of sentences were annotated to contain true serendipitous drug usage mentions. The rest 15,267 sentences were annotated to contain no serendipitous drug usage mentions. This dataset was used throughout the study as the gold standard dataset to train and evaluate various machine learning models.

3.2.5 Feature construction and selection

Feature construction and selection is an important part of data mining analysis, in which the data is processed and presented in a way understandable by machine learning algorithms. The original drug reviews downloaded from WebMD website come with 11 features, including patients' ratings of drug effectiveness, ease of use, overall satisfaction, and the number of people who thought the review is helpful (See Table 5).

In the data-filtering step, we created four more features, which are (1) whether the sentence contains negation, (2) the UMLS semantic types of mentioned medical concepts; (3) the SNOMED CT-based semantic distance between a drug's known indication and the medical concept the user mentioned in a review sentence; (4) the sentiment score of the review sentence.

Prior knowledge in drug discovery and development also tells that some therapeutic areas, such as neurological disorders, bacteria infection, and cancers are more likely to have "dirty" drugs, which bind to many different molecular targets in human body, and tend to have a wide range of effects [37-39]. Therefore, drugs used in those therapeutic areas have higher chance to be repositioned. We manually selected 155 drug usages from those therapeutic areas and used them as binary features, which hopefully capture useful information and improve machine learning predictions of serendipitous drug usages.

We also adopted a commonly used text-mining method, n-gram [40], to generate more textual features. An n-gram is a contiguous sequence of n words from a given text and it captures the pattern about how people use word combination in their communication. We used the *tm* package in R [41] to do this. After the steps of punctuation and stop words removal, word stemming, and rare words pruning, we

Table 5: List of the features constructed for the annotated datasets

Name	Data Type	Source
Original Features obtained from the Patient Forum		
User rating of effectiveness	Numerical	WebMD
User rating of ease of use	Numerical	WebMD
User rating of overall satisfaction	Numerical	WebMD
Number of users who felt the review was helpful	Numerical	WebMD
Number of reviews for the drug	Numerical	WebMD
The day of review	Categorical	WebMD
The hour of review	Categorical	WebMD
User's role (e.g., Patient, Caregiver)	Categorical	WebMD
User's gender	Categorical	WebMD
User's age group	Categorical	WebMD
The time on the drug (e.g., less than 1 month, 1 to 6 months, 6 months to 1 year)	Categorical	WebMD
Additional Features		
Whether the sentence contains negation	Binary	MetaMap
Semantic types of medical concepts mentioned in the sentence	Categorical	MetaMap
Semantic distance between the mentioned medical concept and the drug's known indications in SNOMED CT	Numerical	SNOMED
Sentiment score	Numerical	Deeply Moving
Therapeutic areas (155)	Binary	Self-constructed
N-grams extracted from drug review sentences (177)	Binary	Self-constructed

extracted 3,264 unigrams, 10,064 bigrams, and 5,058 trigrams. For each n-gram, we calculated the information gain [42] to assess its differentiating power between true and false classes in Weka [43]. We excluded n-grams whose information gain equals zero and kept 177 n-grams with positive information gain (namely 64 unigrams, 73 bigrams, and 40 trigrams) as additional textual features. In total, 347 features were constructed for the machine learning classification, as summarized in Table 5.

3.2.6 Data preprocessing

We normalized the data by linearly re-scaling all numerical features to the range of [-1, 1]. Such processing is necessary for support vector machine (SVM) to ensure no features dominate the classification just because of their order of magnitude, as SVM calculates the Euclidean distances between support vectors and the separation hyperplane in high-dimensional space [44]. Then we split the 15,714 annotated sentences into training, validation, and test datasets, according to their post dates. Sixty percent of them, or 9,429 sentences posted between September 18, 2007 and December 07, 2010, were used as the training dataset to build machine learning models. Twenty percent of the data, or 3,142 sentences posted between December 08, 2010 and October 11, 2012 were used as the validation dataset to tune the model parameters. The remaining 20% of data, or 3,143 sentences that were posted between October 12, 2012 and March 26, 2015, were held as the independent test dataset. The proportion of serendipitous drug usages in the three datasets was between 2.0% and 3.2%. This arrangement is essential to pick up the models that could generalize on future and unseen data and minimize the bias led by overfitting.

3.2.7 Machine learning models

We selected three state-of-art machine learning algorithms, namely SVM [45], random forest [46] and AdaBoost.M1 [47] to build the prediction models. The implementation was based on Weka (version 3.7) [43] and LibSVM library [48]. For SVM, we used the radial basis function (RBF) kernel and conducted a grid search to find the optimal parameters including C and gamma (γ). For random forest, we empirically set the number of trees to be 500 and iteratively searched for the optimal value for number of

features. For AdaBoost.M1, we selected the decision tree built by C4.5 algorithm [49] as the weak learner and obtained the optimal value for number of iterations through an iterative search.

As the chance of finding a serendipitous drug usage (positive class) is rare and the vast majority of the drug reviews posted by users do not mention any serendipitous usages (negative class), we were facing an imbalanced dataset problem. Therefore, we used the oversampling technique [50-52] to generate another training dataset where the proportion of positive class was increased from 2.8% to 20%. Afterward, we tried the same machine learning algorithms on the oversampled training dataset, and compared the prediction results side-by-side with those from the original, imbalanced training dataset.

3.2.8 Evaluation

We were cautious about choosing appropriate performance evaluation metrics because of the imbalanced dataset problem. Of commonly used metrics, accuracy is most vulnerable to imbalanced dataset since a model could achieve high accuracy simply by assigning all instances into the majority class. Instead we used a combination of three commonly used metrics, namely precision, recall, and area under the receiver operating characteristic curve (also known as AUC score) [53], to evaluate the performance of various prediction models on the independent test dataset.

In addition, we manually reviewed 10% of instances in the test dataset that were predicted to be serendipitous drug usages and searched through the scientific literature to check if these predictions based purely on machine learning methods can replicate the discoveries from biomedical scientific community. This serves as another verification on

whether machine learning methods alone can potentially predict completely new serendipitous drug usages.

3.3 Results

3.3.1 Model parameters

We used AUC score to tune the model parameters on the validation dataset. In case that the AUC scores of two models were really close, we chose the parameter and model that yielded higher precision. This is because end users (e.g., pharmaceutical scientist) are more sensitive to cases that were predicted to be the under-presented and rare events, which are serendipitous drug usages in this work, when they evaluate the performance of any kind of machine learning based predictive models. For SVM models, the optimal value of gamma (γ), the width of RBF kernel was 0.001 without oversampling and 0.1 with oversampling. The optimal value of *C,* which controls the trade-off between model complexity and ratio of misclassified instances, was equal to 380 without oversampling and 0.1 with oversampling. For random forest models, the number of features decides the maximum number of features used by each decision tree in the forest, which was found to be 243 without oversampling and 84 with oversampling at the best performance on validation dataset. For AdaBoost.M1, the number of iterations specifies how many times the weak learner will be trained to minimize the training error. Its optimal value equaled 36 without oversampling and 58 with oversampling.

3.3.2 Model performance metrics

We evaluated the performance of six prediction models, namely SVM, random forest and AdaBoost.M1 with and without oversampling, on independent test dataset. The results were summarized in Table 6. The highest AUC score (0.937) was achieved from

the AdaBoost.M1 model, whereas the lowest score (0.893) was from the SVM with oversampling. On the whole, AUC scores for all models were higher than 0.89, demonstrating the promise of machine learning models for identifying serendipitous drug usages from patient forums.

The precision of random forest and AdaBoost.M1 models with and without oversampling, and the SVM model without oversampling were between 0.758 and 0.857, with the highest precision achieved on the random forest model without oversampling. However, the precision for the SVM model with oversampling was 0.474, which was significantly lower than the other models. The recall of all models was less than 0.50. This means more than 50% of serendipitous usages were not identified. Obtaining either low recall or low precision remains a common challenge for making predictions from extremely imbalanced datasets like ours [50]. In many cases, it becomes a compromise depending on the application and the users' need. In our experiment, after we increased the proportion of the positive class to 20% by oversampling, the recall of SVM and random forest models increased slightly; but the precision and the AUC score decreased. Oversampling seemed ineffective on AdaBoost.M1 models. The AUC score, precision

Table 6: Model performance in terms of AUC score, precision, and recall

Model	Test dataset			10-fold cross validation		
	AUC	Precision	Recall	AUC	Precision	Recall
SVM	0.900	0.758	0.397	0.926	0.817	0.539
SVM - Oversampling	0.893	0.474	0.429	0.932	0.470	0.620
Random Forest	0.926	0.857	0.381	0.935	0.840	0.506
Random Forest - Oversampling	0.915	0.781	0.397	0.944	0.866	0.530
AdaBoost.M1	0.937	0.811	0.476	0.949	0.791	0.575
AdaBoost.M1 - Oversampling	0.934	0.800	0.444	0.950	0.769	0.559

and recall for AdaBoost.M1 with oversampling all decreased, compared to the metrics on AdaBoost.M1 models without oversampling.

In order to compare our results directly with some other drug-repositioning studies, we also conducted a 10-fold cross validation by combining training, validation and testing datasets together. It seems that both recall and AUC scores from the 10-fold cross validation were better than what were observed on the independent test set. Our AUC scores were close to the same scores reported by the drug-repositioning algorithm of PREDICT [11], which were also from a 10-fold cross validation.

3.3.3 Review of predictions

For the 10% of instances in the test dataset that were predicted to be serendipitous drug usages, we conducted a search in literature and clinical trials to provide a closer verification of our prediction models. Table 7 summarizes the analysis. We also presented the condensed evidences in literature and/or clinical trial below, for each instance.

Metformin and obesity: A patient reported weight loss while taking metformin, a type 2 diabetes drug. Actually in the past two decades, metformin's effectiveness and safety for treating obesity in adult and child patients have been clinically examined in dozens of clinical trials and meta-analyses studies with promising results [54-58]. According to a literature review published in 2016 [54], one possible explanation is that metformin could increase the body’s insulin sensitivity, which helps obese patients (who typically develop resistance to insulin) to reduce their craving for carbohydrates and to reduce the glucose stored in their adipose tissue. Other explanations include that metformin may enhance energy metabolism by accelerating the phosphorylation of the

AMP-activated protein kinase system, or it may cause appetite loss by correcting the sensitivity and resistance of leptin [54].

Painkiller and depression: When tramadol was taken for back pain, a patient found it also helpful with his depression and anxiety. Tramadol is an opioid medication, which have been long used for the psychotherapeutic benefits [59]. Tetsunaga *et al.* have demonstrated tramadol's efficacy in reducing depression levels among lower back pain patients with depression in an 8-week clinical trial. The self-reported depression scale of patients in the tramadol group was 6.5 points lower than the control group [60]. Similarly the combinatory therapy of acetaminophen and oxycodone, another painkiller, was reported to have antidepressant effect too [61].

Bupropion and obesity: In the specific comment, the patient reported that Bupropion, an anti-depressant, helped him to lose weight. The weight loss effect of bupropion might be attributed to increased dopamine concentration in the brain, which leads to suppressed appetite and reduced food intake [62]. This serendipitous drug usage was also supported by several clinical trials [63-65].

Ondansetron and irritable bowel syndrome with diarrhea: Ondansetron is a medication for nausea and vomiting. Sometimes it causes the side effect of constipation in patients. Interestingly, this patient also had irritable bowel syndrome with diarrhea and thus ondansetron helped to regulate that. This serendipitous usage actually highlights the justification of personalized medicine and has been tested in a recent clinical trial [66].

Table 7: Examples of serendipitous drug usages predicted by the models

True positive examples										
Drug	**Known indications**	**Serendipitous usage**	**Example**	**SVM**	**SVM-Oversampling**	**Random Forest**	**RF-Oversampling***	**AdaBoost**	**Ada-Oversampling***	**Literature evidence**
Metformin	Type 2 Diabetes Mellitus, Polycystic Ovary Syndrome, etc.	Obesity	I feel AWFUL most of the day, but the ***weight loss*** is great.	x	x	x	x	x	x	[54 - 58]
Tramadol	Pain	Depression, anxiety	It also has helped with my ***depression*** and ***anxiety***.	x	x			x	x	[60]
Acetaminophen & oxycodone	Pain	Depression	While taking for pain I have also found it relieves my major ***depression*** and actually gives me the energy and a clear mind to do things.	x	x	x		x		[61]
Bupropion	Depression, attention deficit & hyperactivity disorder	Obesity	I had energy and experienced needed ***weight loss*** and was very pleased, as I did not do well on SSRI or SNRIs.	x	x		x	x	x	[62 - 65]
Ondansetron	Vomiting	Irritable bowel syndrome with diarrhea	A lot of people have trouble with the constipation that comes with it, but since I have ***IBS-D*** (irritable bowel syndrome with diarrhea), it has actually regulated me .					x	x	[66]
Desvenlafaxine	Depression	Lack of energy	I have had a very positive mood and ***energy*** change, while also experiencing much less anxiety.	x	x	x	x	x		
False positive examples										
5-HTP	Anxiety, depression	Thyroid Diseases, Obesity	i have ***Hoshimitos thyroid disease*** and keeping stress levels down is extremely important for many reasons but also for ***weight loss***.		x		x			
Cyclobenzaprine	Muscle spasm	Pain	While taking this medication for neck stiffness and ***pain***; I discovered it also helped with other muscle spasms.		x					

Desvenlafaxin and lack of energy: In the last case, anti-depressant desvenlafaxine was reported to boost energy. Strictly speaking, lack of energy is not a disease but a symptom. With limited information on the patient's physical and psychological conditions before and after medication, it remains unclear whether the energy boost effect was due to changes in the neural system or was purely a natural reflection of more positive moods after the patient took the anti-depressant medicine. We did not find any scientific literature discussing the energy boost effect of desvenlafaxine. So this case could represent either a new serendipitous drug use or a promiscuous drug usage.

False positive predictions: Besides the true positive examples, we also found two cases where some of our models made false positive predictions due to difficult language expression and terminology flaw (See Table 7). The first example is 5-HTP, an over-the-counter drug for anxiety and depression. One patient commented that stress relief brought by this drug was important to her Hashimito's thyroid disease and weight loss. Although Hashimoto's disease and weight loss were mentioned, the patient did not imply the 5-HTP can treat Hashimoto’s disease or control weight. But SVM and random forest models with over-sampling became confused by the subtle semantic difference. In the second case, a patient taking cyclobenzaprine for neck stiffness and pain said the drug also helped with other muscle spasms. Pain, neck stiffness and muscle spasms are really close medical concepts. We found that this false positive prediction was actually due to imperfect terminology mapping.

3.4 Discussion

In this very first effort to identify serendipitous drug usages from online patient forum, we designed an entire computational pipeline. This feasibility study enabled us to

thoroughly examine the technical hurdles in the entire workflow and answer the question if patient-reported medication outcome data on social media is worthwhile to explore for drug repositioning research. The best-performing model was built from AdaBoost.M1 method without oversampling, which had precision equal to 0.811, recall equal to 0.476 and AUC score equal to 0.937 on independent test data. The 10-fold cross validation results are also comparable to existing drug-repositioning method [11]. Therefore, we are more confident in applying machine learning methods to identify serendipitous drug usages from online patient forum data. More specifically, we have addressed the following tasks in this work:

Previously, there was no annotated social media dataset available for the purpose of identifying serendipitous drug usages. We spent a considerable amount of time and effort to collect, filter and annotate 15,714 drug review sentences from the WebMD patient forum site. This annotated dataset is comprehensive enough to cover not only easy instances, but also challenging ones for machine learning prediction, as shown in Table 7. It can be used as the gold standard for current and future research in drug repositioning.

In addition, the drug reviews posted on patient forum are unstructured and in an informal human language, which is prevalent with typographic errors and chat slangs. These reviews need to be transformed to a representation of feature vectors before machine learning algorithms could comprehend. We used patients' demographic information, ratings of drug effectiveness, ease of use, and overall satisfaction from the patient forum. We calculated negation, the sentiment score for each sentence, and the semantic similarity between the unexpected medication outcome mentioned in a review sentence and the known drug indications based on SNOMED CT. We also leveraged our

known knowledge on dirty drugs, and extracted informative n-gram features based on information gain. The results from this feasibility study showed that these features are useful to predict serendipitous drug usages. For example, dirty drugs for neurological conditions did show up predominantly in the results. But these features seemed not sufficient to predict all serendipitous drug usages correctly. As shown in the false positive examples of Table 7, the n-grams such as *also*, *also help*, and *also for* were often associated with true serendipitous drug usages, but could occur in false positive cases too. Current medical terminology mapping tools (i.e., MetaMap) could be the performance-limiting step in cases like *pain* and *muscle spasm*, despite the close connection of these two concepts from the perspective of medicine. Future efforts are needed to improve terminology mapping accuracy, for example, using more sophisticated terminology mapping tools such as DNorm [67].

Thirdly, the data are extremely imbalanced between two classes (2.8% vs. 97.2%) because serendipitous drug usages are rare events by nature. Such imbalance inevitably impedes the performance of machine learning algorithms. We tried to increase the proportion of serendipitous usages in the training dataset to 20%, using the random oversampling method [50]. We have also tried two other methods, namely synthetic minority over-sampling technique [68] and under-sampling [52], but their performance was inferior to that of random oversampling and not shown here. More robust machine learning algorithms that are less sensitive to imbalanced data or robust sampling methods will be desirable to further improve serendipitous drug usage predictions.

Last but not least, we acknowledge that as an emerging data source, online patient forums have limitations too. Many patients who write drug reviews online lack of basic

medical knowledge. Their description of the medication experience can be ambiguous, hyperbolic or inaccurate. Also, important contextual information, such as co-prescribed drugs, may be missed in the review. Without a comparison between an experiment group and a control group, serendipitous drug usages extracted from patient forums need to be further verified for drug repositioning opportunities by integrating with existing data sources, such as EHR and scientific literature.

CHAPTER 4: DEEP LEARNING FOR PREDICTING SERENDIPITOUS DRUG USAGES IN SOCIAL MEDIA TEXT

4.1 Background

In previous chapters, we surveyed patient-reported medication outcome information on social media and found that patients do report serendipitous drug usages on social media, which can be important clues to generate and validate drug repositioning hypotheses (Chapter 2). To systematically identify these mentions in patient forum data, we curated a gold-standard dataset based on the filtered drug reviews from WebMD and built a computational pipeline of machine learning and text mining modules to predict serendipitous drug usages (Chapter 3). Our models achieved AUC scores that are comparable to the existing drug repositioning methods [11]. Many instances predicted to be serendipitous drug usages are also supported by the scientific literature.

Recently, deep learning methods became popular in the data mining community and have achieved great progresses in computer vision research. This encouraged some researchers to apply deep learning models that was designed for vision analysis (e.g., Convolutional Neural Network) to text classification tasks. These works often redesigned deep learning models to fit on several well-known annotated text mining datasets, such as rotten tomato movie reviews, news articles, and product reviews, and usually reported that deep learning models outperformed traditional machine learning models such as SVM and logistic regression [69, 70]. However, none of the datasets used in these studies are extremely imbalanced between classes – which often happens in social media data. In this chapter, we investigate deep learning methods in the context of identifying serendipitous drug usages in patient forum data. We introduced word embedding as a

new method to construct machine learning features from social media text and then designed four deep learning models to predict true mentions of serendipitous drug usages.

4.2 Method

4.2.1 Feature construction using word embedding

A word embedding is a matrix that represents words as dense vectors in a high-dimensional vector space, as illustrated in Figure 5. It was extracted from a neural network that was trained on large unannotated text corpus [71-73]. Studies shown that in the vector space of a word embedding, words with syntax and semantic relations tend to be close to each other [72, 74, 75].

```
emb = {
        'drug':  [-0.137818,-0.131454, 0.063686,...,-0.119478],
        'also' :  [ 0.123257,-0.013544, 0.030601,..., 0.039448],
        ...
        'help':  [ 0.047297, 0.119153, -0.001935,...,-0.127605]
        }
```

Figure 5: A word embedding implemented in Python

Among models to generate word embedding [71, 72, 75], Word2Vec is very popular in recent text mining research and competitions [76, 77]. It is a shadow neural network of an input layer, a projection layer, and an output layer [72]. Word2Vec has two forms, namely continuous bag-of-words (CBOW) and skip-gram (Figure 6). The CBOW takes a number of words as input (or condition) to predict the probability for target word w_t to appear among these words, the skip-gram takes the word w_t as input to predict the probability for other words to surround it. The context window parameter defines how many words before or after the target word are considered as 'appear among' or

'surrounded by'. The projection layer contains *d* neurons, which defines the dimension for word vectors. The word embedding consists of the weights for connections between the words and projection neurons. While training the model, these weights were adjusted by backpropagation according to actual probabilities of words observed in the text corpus.

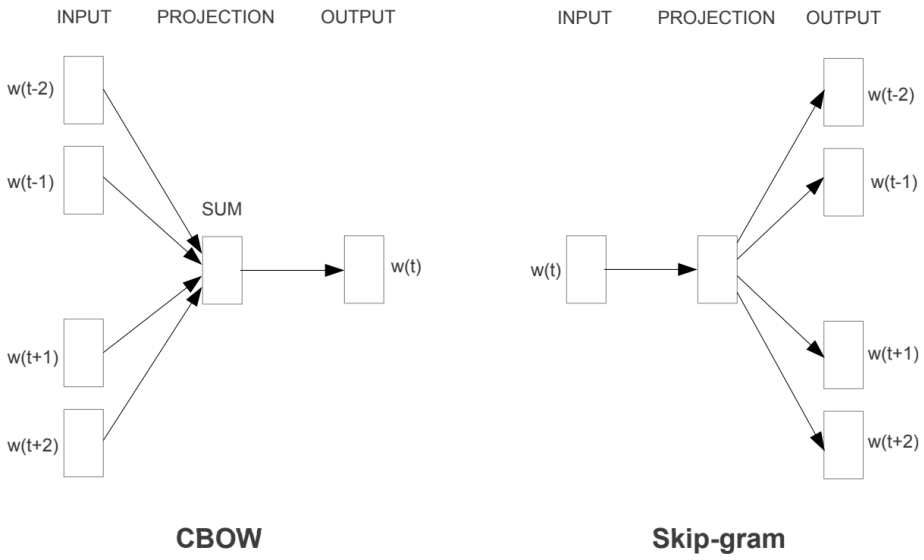

Figure 6: Word2Vec models – from Mikolov *et al.*(2013) [72]

In this study, we implemented a word embedding from all drug review sentences that we collected from WebMD, using the Word2Vec model from the Gensim Python library [78]. We removed non-English characters, converted all letters to lower case, and stemmed words to its basic form. We chose CBOW as it is suggested by Gensim for most text mining tasks. Of model configuration, we set the dimension of word vectors – *d* to be 200 and specified the context window size to be 50, because over 99% of sentences in the WebMD corpus are shorter than 50 words (Table 8). For other options, we adopted the default configuration of the Gensim library. The resulting WebMD word embedding contains 67,659 word vectors with the dimension of 200.

Table 8: Number of words per sentence

Quantile	5%	25%	50%	75%	95%	99%
Number of words per sentence	5	9	13	19	33	50

After obtaining the word embedding, the next step is constructing machine learning features from sentences. There are three common methods to compose sentence vectors from the word embedding. The first method is based on vector aggregation. It sums up the vectors of words in a sentence and optionally divides the aggregated vector by the length (number of words) of the sentence [77]. The sentence vector has the same dimensionality as each word vector, making it convenient to use with most classification models. The second method adopts clustering [79]. Since semantically related words in a word embedding tend to close to each other, they can be grouped to k clusters using algorithms such as K-Means. Then, we can represent a sentence as a vector of k dimensions, with the value in each dimension equals to the count of words belonging to a cluster. The clustering method encodes the semantic information in a sentence, but the computational cost of clustering is expensive. The third method stacks word vectors to a sentence matrix. For a sentence *S* of *i* words w_1, w_2, …, w_i, it stacks word vectors v_{w1}, v_{w2}, …, v_{wi}, each has *d* dimensions, to form a *i* x *d* dimension matrix (Figure 7). This approach appeared in recent text mining that utilized deep learning models [70, 80-82].

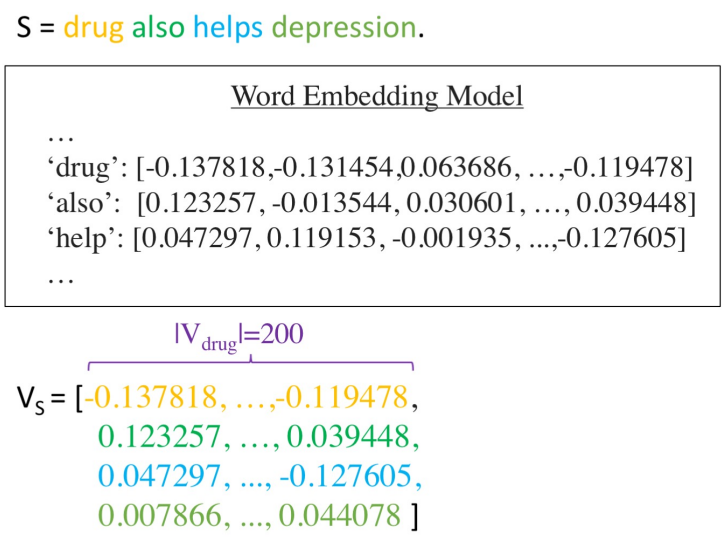

Figure 7: Sentence as concatenation of word vectors

Comparing three word embedding based feature construction methods, the aggregation and clustering methods construct dense sentence vectors with fixed size but lose the sequential pattern between words; the stacking method preserves the sequential pattern, but the sentence matrixes are in various sizes. In this study, we chose the stacking approach because certain word sequences (e.g., 'it also helps') are important for identifying true serendipitous drug usage mentions. To unify the shape of matrixes, we limited the number of word vectors to be 50 for each sentence by padding zeros for shorter sentences and trimming excessive words from longer sentences.

4.2.2 Deep Learning Models

In recent years, deep learning models achieved remarkable results in data mining tasks such as image classification and self-driving vehicle [83-85]. Among deep learning models, convolutional neural networks (CNN) is one of the most successful models. CNN utilizes convolution filters to learn patterns in the data at different levels of abstraction. It could handle different transformations of the data to generalize. Although

CNN is popular in computer vision research, emerging research shown that it also achieved outstanding performance in text classification tasks [80, 81, 85]. Inspiring by promising results of these studies, we advanced to apply CNN models in identifying serendipitous drug usages from patient health forums. We began with the CNN model from Kim (2014) [80] and revised the architecture incrementally to explore a design adaptive to our text mining task.

CNN: The CNN static model in Kim (2014) [80] transformed the input sentence to a matrix using the stacking approach that we mentioned in the feature construction section. The model used the word embedding from Google to generate feature vector for texts. It contained paralleled convolution filters of three different sizes, followed by max pooling filters, and concatenated outputs to a fully connected layer of neurons to make the prediction.

Our first model adopted these designs with our WebMD word embedding features (Figure 8). The convolution filters were trained to protrude informative patterns in a sub area of the input data. Each filter is designed to enhance part of the input signal that matches a specific pattern by raising some values in the input vector. As the number of filters increase, each filter will focus on a more specific pattern. The kernel size (k) of filter determines the magnitude of the sub area – in the case of text, that is the number of continuous words in a sentence. By mixing convolution filters of three continuous sizes (k-1, k, and k+1), the network can learn patterns in the sentence at three different scales. This mechanism is similar to combining n-gram features of different scales (e.g., unigram, bigram, and trigram) in feature construction. For the sake of convenience, we created same number of filters for each kernel size. The max pooling filter scan through outputs

of convolution filters and preserve only the max value in each area. This operation washes out information that is less relevant to the classification task and reduces the dimensionality of features extracted by convolution filters. The output was further processed by one layer of fully-connected neurons ('Dense 0') before submitted to a single neuron ('Prediction') to make the prediction. Additionally, we inserted one Dropout layer ('Dropout 0') after embedding matrix and another one ('Dropout 1') before the prediction neuron. Dropout layer randomly intercepts output of previous neurons to next layer to prevent overfitting. We referred this model as the CNN model and fitted it on our annotated dataset.

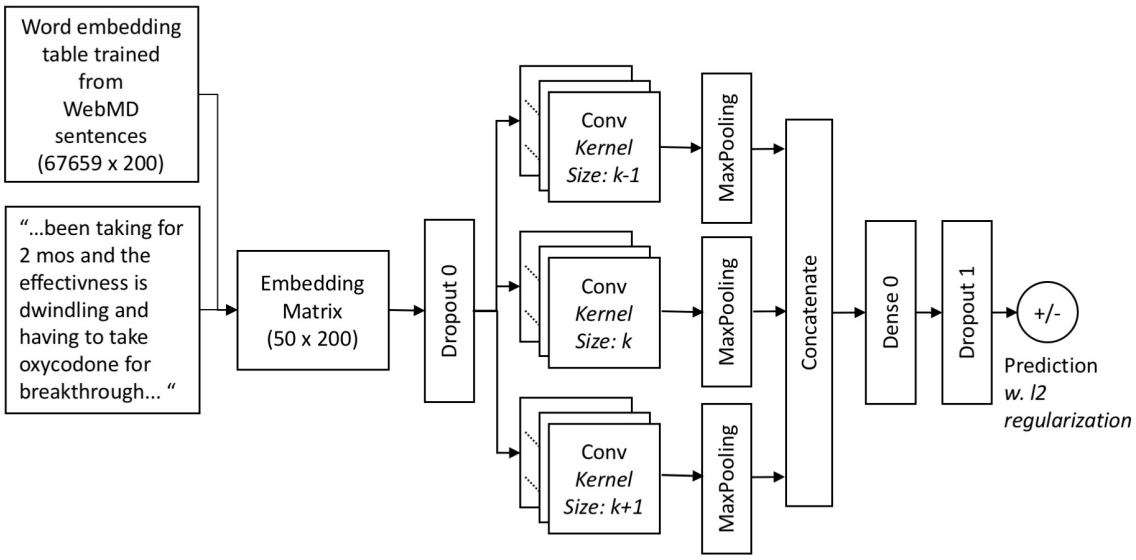

Figure 8: The CNN model for text classification – based on Kim (2014) [80]

CNN with non-text embedding: The CNN model takes only sentences as inputs, but our annotated dataset also includes non-text features extracted from original drug reviews, drug knowledge, and information filtering tools. To utilize all available information for making predictions, we added these non-text features to the CNN model

by appending them as an additional word vector to the sentence embedding matrix in the CNN model (Figure 9). This design applies convolution filters to both text and non-text features. To keep the non-text feature vector in the alignment with the sentence matrix, we transformed all nominal type features to the binary type, resulting a feature vector of 195 dimension. Then we padded five zeros at the end of the vector to match the dimensionality of the sentence embedding matrix. The other model components remain unchanged.

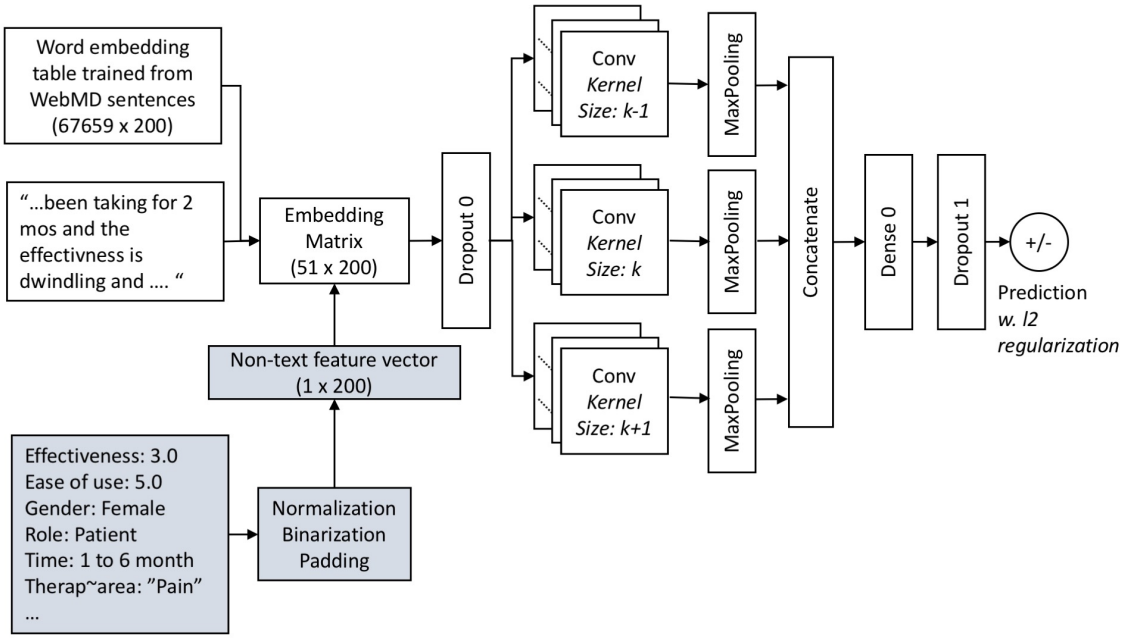

Figure 9: CNN with non-text feature embedding model

Paralleled CNN and Fully Connected Neural Network (FCN): Our third model applied different types of neural network to text and non-text features. For text features, we adopted the architecture of the CNN model but removed the prediction layer. For non-text features, we designed a neural network of three fully connected layers ('Dense 1-3'), with each layer containing half of neurons of the previous one. At the end, we combined

outputs from text features and non-text features together to an additional layer of fully connected neurons (‘Dense 4’) and another dropout layer (‘Dropout 2’) before making the prediction (Figure 10).

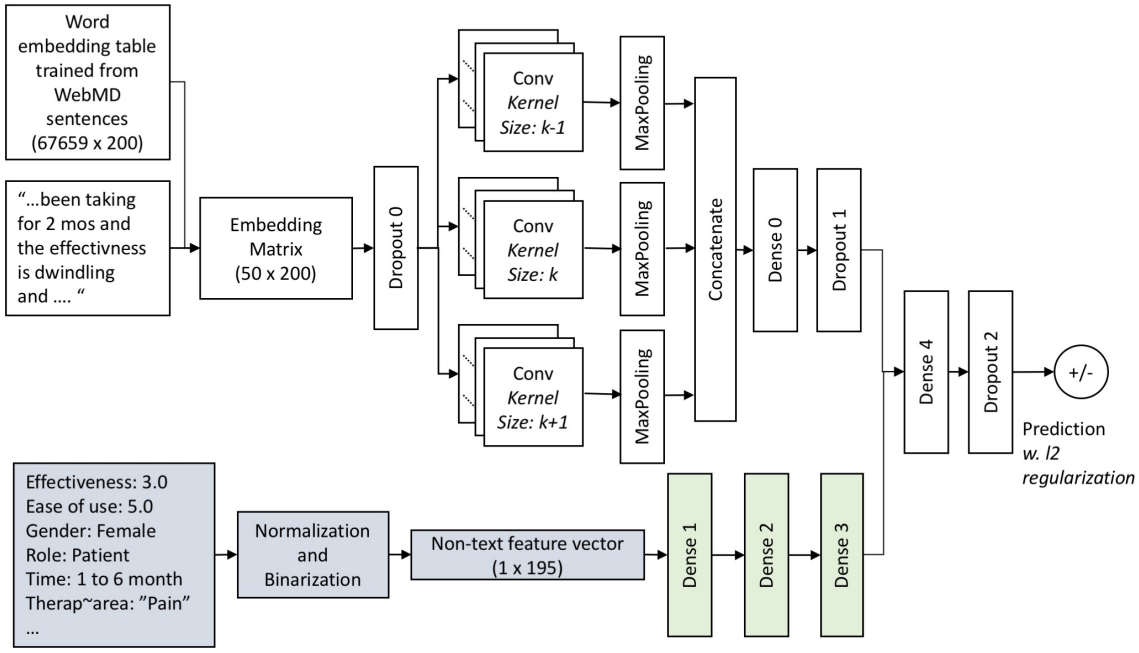

Figure 10: Paralleled CNN and FCN model

Paralleled CNN-LSTM and FCN: By far, we utilized convolution filters to extract signals from social media texts. Though convolution filters are good at dealing with data in the matrix or grid representation, they can only capture sequential patterns in a local area and inevitably ignore long-range dependencies between words of a sentence [86]. To solve this problem, Hochreiter *et al.* introduced the Long Short Term Memory networks (LSTM) [87-89]. LSTM is a special type of recurrent neural network. It contains a chain of repeating neural networks units. The number of units determines the length of sequence – the number of words in our case – for the LSTM network to process. Whiling processing sequential data, LSTM network leverages four information gates to

decide how much new information to add to and old information to remove from the information flow [88]. LSTM emerged in recent text mining research works [70, 90]. In the fourth model, we adopted Zhou *et al.* (2015) approach [70] by adding LSTM behind the max pooling filters of the third model (Figure 11). In the new model, convolution and max pooling filters extract the most discriminative local patterns and continuously feed the signals to LSTM, which concentrates on learning the sequential patterns in the information flow.

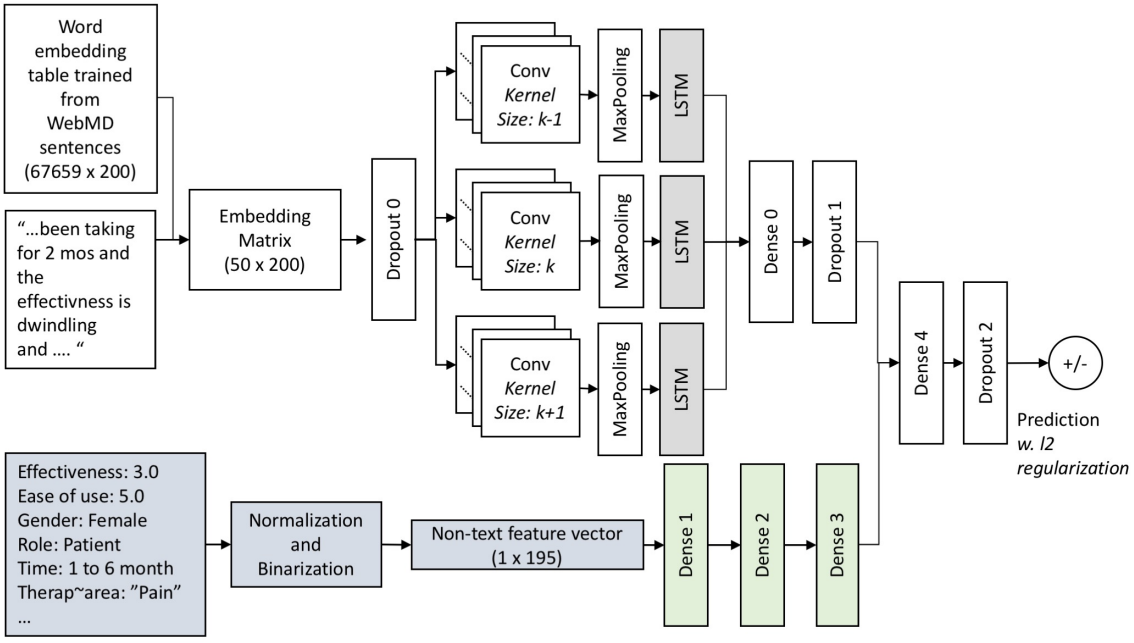

Figure 11: Parallel CNN-LSTM and FCN model

4.2.3 Model Implementation

Platform: We implemented our deep learning models on a Google Cloud virtual machine that equips six CPUs, 20 GB memory, and one slice of NVIDIA Tesla K80 GPU. The software environment consists of Ubuntu 16 Linux System, Python 3.6.2, Keras

2.0.8 – a famous deep learning library for Python [91], and Tensorflow 1.3.0 math library [92] for high efficiency neural network computing.

Model configuration: We chose rectified linear unit (ReLU) [93] as the activation function for all neural network layers except the prediction layer. ReLU computes the function of $relu(x)=max(x,0)$. It does not saturate (the gradients do not diminish when x approaches positive infinity) and is less computationally expensive than functions that involve the exponential calculation, making it popular in recent deep learning research. For prediction layer, we chose the sigmoid function: $sigmoid(z)=1/(1+e^{-z})$, which takes a real value input to an output in the range from 0 to 1 [94]. We selected Adam as the kernel optimizer [95] and binary cross entropy as the loss function [96]. We split 15,714 instances in the annotated dataset into training, validation, and testing parts, in the same way we mentioned in the Chapter 3.

Hyper parameter tuning: Hyper parameters are configurations that impacts the model's behavior in machine learning tasks. Their values cannot be directly estimated from data [97]. They are often specified by the user based on heuristics techniques such as rules of thumb and conventions. In practice, people often tune hyper parameters to improve the model's performance for each specific machine learning problem. For our deep learning models, the hyper parameters we tuned include the kernel size of convolution filters, the number of convolution filters, the size of pooling window for Max Pooling filters, the number of neurons for each Dense layer, the drop ratio for each Dropout layer, the constant parameter of the *l2* kernel regularizer [98] for the prediction neuron, and the number of units for the LSTM layer. We also searched for the best method to initialize the weights of neural network, among six commonly used initializers:

random uniform, random normal, Xavier uniform, Xavier normal, He uniform, and He normal [94, 99]. Moreover, neural networks are sensitive to imbalanced data. Keras provides a cost sensitive learning solution by allowing us to specify the importance of each class in the class weights configuration. To fully leverage this feature, we treat class weights as an additional hyper parameter to tune.

We fit models on training dataset with different hyper parameter sets and track these models' performance on the validation dataset. We used Hyperas [100], a Keras wrapper for famous parameter tuning library of Hyperopt [101]. Unlike grid search, Hyperas does not exhaustively search the entire hyper parameter space for the optimal set. Instead, it leverages search algorithms such as random search and Tree of Parzen Estimators [102] and advanced parallel computing methods to partially search the parameter space for a relatively good parameter setting. It is widely adopted in deep learning research and application because the search space for hyper parameters are often too big for grid search method to complete in reasonable amount of time. We used Precision-Recall score as the optimization target for Hyperas. The score equals to the area under the Precision-Recall curve, which can be plotted based on the model's precision and recall at different classification thresholds. To our knowledge, recent deep learning research often used the accuracy score as the target to optimize model parameters [83, 84, 90]. However, the datasets used by these studies are not as imbalanced as ours. For highly imbalanced datasets, Precision-Recall score is more informative on model performance than AUC, accuracy, precision, and recall [103].

Evaluation: We adopted AUC, precision, and recall to assess the performance of our models on testing dataset and compared models in terms of machine learning performance metrics and model complexity.

4.3 Results

4.3.1 Hyper parameters

Using Hyperas and the validation dataset, we tuned hyper parameters for each model and list the results in Table 9. We cautiously warn that the listed values can be sub-optimal because Hyperas only partially search the parameter space and determines the optimal value by the optimization algorithm and the loss function. We note that kernel size of convolution filters is higher for CNN and CNN w. non-text embedding models than the other two models. This reflects associations between the function of convolution filter, the characteristic of input data, and the structure of model. In deep learning models, convolution filters are used primarily to extract informative patterns from local area of a matrix. Its kernel size determines the scale of the local area, which is the continuous number words in our case. Since 75% of sentences in WebMD dataset contain less than 19 words and 50% less than 13 words (Table 8). For CNN and CNN w. non-text embedding models, the optimal kernel sizes need to be larger (11 to 16 words) to cover sufficient number of words in each sentence. The kernel size for Paralleled CNN-LSTM and FCN model is the smallest. This might attribute to the existence of LSTM layers. LSTM can learn patterns from word sequence. With LSTM to recognize patterns in the length of 40 words, CNN filters can focus on patterns in much smaller (5 to 7 words) local area.

Table 9: Hyper parameters by model

		CNN	CNN w. non-text embedding	Paralleled CNN and FCN	Paralleled CNN-LSTM and FCN
Conv kernel sizes (*k-1, k, k+1*)		11,12,13	14,15,16	7,8,9	5,6,7
# of Conv filters per kernel size		32	64	256	64
# of units per LSTM		--	--	--	40
Size of MaxPooling filter		4	5	5	2
# of neurons	Dense_0	128	32	32	256
	Dense_1	--	--	256	256
	Dense_2*	--	--	128	128
	Dense_3*	--	--	64	64
	Dense_4	--	--	128	128
Dropout rate	Dropout_0	0.8639	0.0875	0.5300	0.5045
	Dropout_1	0.5319	0.4911	0.1961	0.4131
	Dropout_2	--	--	0.0625	0.1119
l2 constant		9.3467	6.1523	3.0267	2.2294
Initialization method		He uniform	Random uniform	He uniform	He normal
Class weights (neg : pos)		1:29.8666	1:16.5773	1:25.7443	1:14.1905

***** The number of neurons for Dense_2 and Dense_3 are designed to be half and quarter of Dense_1

Additionally, values for the class weights parameter reflect the difference between models in the strength of imbalance correction. Comparing CNN and CNN w. non-text embedding models, the former demands the data to be more balanced to achieve its best performance. We may indicate the introduction of non-text feature improve the tolerance of CNN to imbalanced data. Similarly, we may also suggest adding LSTM improved the tolerance of Paralleled CNN and FCN model.

We are curious on the how much did hyper parameter tuning impact the performance of models. Table 10 lists the minimum, average, and maximum AUC, Precision, and Recall of models on validation dataset with different sets of hyper parameters. The wide spread between best and worst parameter sets in all machine

learning performance metrics indicate the need of tuning hyper parameters for deep learning models.

Table 10: Impact of hyper parameters

		CNN	CNN w. non-text embedding	Paralleled CNN and FCN	Paralleled CNN-LSTM and FCN
AUC	Min	0.556	0.230	0.695	0.654
	Avg	0.843	0.844	0.827	0.803
	Max	0.908	0.897	0.908	0.894
Precision	Min	0	0	0	0
	Avg	0.151	0.454	0.501	0.421
	Max	0.333	1.0	1.0	0.769
Recall	Min	0	0	0	0
	Avg	0.239	0.142	0.331	0.363
	Max	0.746	0.524	0.476	0.730

4.3.3 Model evaluation

We evaluated deep learning models in terms of AUC, Precision, Recall, and Precision-Recall scores on the test dataset and set true serendipitous usage as the positive class. Table 11 summaries the results. The highest AUC score (0.919) is from CNN model. The lowest score (0.815) is from the Paralleled CNN and FCN model. The precision of deep learning models spread widely between 0.156 to 0.735, with Paralleled CNN and FCN model achieving the highest precision and CNN model the lowest. The recall of deep learning models ranges from 0.317 to 0.683. CNN and Paralleled CNN-LSTM and FCN models are better than CNN w. non-text embedding and CNN. Model performance metrics indicate introducing non-text features can greatly improve the precision of models and relieve model from overfitting issue. Additionally, FCN are more suitable to process non-text features for our study. This might because the vector of non-text features is much sparser than the word embedding vector. The stacked layers of FCN

with decreasing number of neurons actually reduce the dimensionality and sparsity of the vector, which usually improve machine learning results.

Table 11: Model performance on testing dataset

	AUC	Precision	Recall
CNN	**0.919**	0.156	**0.683**
CNN w. non-text embedding	0.866	0.606	0.317
Paralleled CNN and FCN	0.815	**0.735**	0.397
Paralleled CNN-LSTM and FCN	0.865	0.659	0.460

Besides performance matrices, complexity is also important to select deep learning models. Table 12 compare the input dimensions, number of three types of neural network filters, and total number of weights for each model. The number of weights can quantify the complexity of deep learning models, as more complex models have more connections between neurons and require more weights to learn in the training process. Among our models, CNN and Paralleled CNN-LSTM and FCN model have significantly less number of weights to train than the other two models. Considering both performance metrics and complexity, we conclude Paralleled CNN-LSTM and FCN model is the best of the four.

Comparing to models that we explored in Chapter 3, none of our deep learning models exceed the AdaBoost.M1 model in any of AUC, Precision, and Recall. While recall of Paralleled CNN-LSTM and FCN model (0.46) is very close to AdaBoost.M1 (0.476), precision is lower (0.659 vs. 0.811). Though it is difficult to determine the reason behind that, we cautiously suggest two factors to consider. First, deep learning models are much more complex by the number of trainable weights (Table 12). Weights determine the importance of each connection between neurons of a neural network. Their role is

equivalent to the support vectors of SVM and the trees in the random forest. Typical deep learning models connect hundreds to thousands of neurons, causing the number of weights far proceeds the number of equivalent components in other models. The large number of weights requires big volume of data to train. In our experiment, although we had 15,714 annotated sentences, we only have 447 in positive class, which can be insufficient for CNN and LSTM filters. We expect deep learning models to exceed current performance should more annotated data – especially positive cases – become available. Secondly, we used grid search – a complete search – to find optimal hyper parameter sets for SVM, random forest, and AdaBoost.M1 models. A complete search in parameter space has greater chance to reach optimal point than partial search approach that Hyperas utlized. In other words, there can exist another set of hyper parameters, with which the deep learning models may perform better than AdaBoost.M1. However, the large number of hyper parameters makes grid-search too computationally expensive for deep learning models.

Table 12: Model complexity

	CNN	**CNN w. non-text embedding**	**Paralleled CNN and FCN**	**Paralleled CNN-LSTM and FCN**
Input dimensions	10,000	10,200	10,195	10,195
# of CNN filters	96	192	768	192
# of FCN filters	128	32	480	480
# of LSTM filters	--	--	--	120
# of weights to train	345,889	658,113	1,532,833	446,561

4.4 Discussion

The success of deep learning methods in computer vision research attracted researchers and data scientists to apply models such as CNN and CNN-LSTM models in

recent text mining research [69, 70, 80-82]. Though deep learning models were reported to exceed baseline models such as SVM and random forest in several recent publications [70, 81, 82], these comparisons were often based on annotated text corpus such as sentiments of movie reviews and categories of news article. Comparing to drug review comments on social media, these data are well balanced and have relatively less domain specific context. Moreover, detailed explanation of how convolution filters work on text data is missing in these works. In this dissertation, we redesigned CNN and CNN-LSTM models to include context information associated with medication outcomes and discussed the impacts of different model components and configurations on text data. We evaluated deep learning models of different structures in identifying serendipitous drug usages in social media text. The results indicate that domain specific context – the non-text features in our case – are important to prevent model from overfitting. Hyper parameters, amount of annotated data and the balance between classes are important to the performance of deep learning models. However, these findings and indications need to be justified by following limitations. First, the total number of sentences in positive class might be insufficient to fully leverage the power of deep learning models. Secondly, we adopted some model designs and configurations directly from previous research without rigorously testing them with alternatives. Additionally, we did not apply our models on commonly used text corpus to verify if indications we made on our data also stand elsewhere.

CHAPTER 5: AN OPEN SOURCE SOFTWARE APPLICATION FOR MINING SERENDIPITOUS DRUG USAGES IN SOCIAL MEDIA TEXT

5.1 Background

In Chapter 3 and 4, we investigated state-of-the-art NLP and machine learning methods to identify serendipitous drug usages mentioned in social media text. The promising results encouraged us to share our prediction models with more users who are interested in this topic. In the research, we leveraged NLP and machine learning software packages, and medical ontologies, including Stanford CoreNLP [32] for sentence boundary detection, text tokenization, and sentiment analysis, MetaMap [33] for recognizing and normalizing disease and symptom mentions in the text, SNOMED CT [34] for measuring the semantic similarity between different medical concepts, and Weka [43], Scikit-learn [104], and Keras [91] libraries for various machine learning algorithms. In our knowledge discovery research work, these tools were integrated in a manually-supervised, step-by-step, and ad-hoc way to provide best agility to frequent changes of research requirements, it is difficult for people with limited knowledge and experience of these tools and our project to use. Therefore, we implemented the *Serendipity* – an open source software application that ensembles the NLP and machine learning methods we had explored to detect serendipitous drug usages from social media text.

To our knowledge, companies such as Treato [105] have created social media text analysis tools for healthcare stakeholders. However, these tools focus on medication effectiveness and side effects rather than serendipitous usages. They provide information retrieval and analytics functions but keep technical details as business secrets. Their commercial licenses also restrict how researchers and software developers could use the

tool. We hope by developing this open-source software application, social media based serendipitous drug usage detection could receive more attentions from drug discovery and development and informatics communities. In this chapter, we document the design and implementation of *Serendipity* at multiple abstraction levels.

5.2 Design overview

We set the following principles in designing: (1) the user should be able to use the application on drug review texts collected from all kinds of social media website; (2) the application should provide functions covering major NLP and machine learning methods available in this dissertation; (3) the application only uses programming language, libraries, and third party software that is distributed under the open source license; (4) the deployment of application should minimize the computational power required for the end users' computer.

We adopted the design pattern of Model-View-Controller (MVC) [106] and implemented the software as a client-server architecture application using Flask – a light weight Python framework for web development [107]. The view layer provides HTML pages and RESTful web services [108] to interact with users, the model layer provides NLP and machine learning functions, and the controller layer coordinates information flow between the view and model layers.

5.3 User and system interaction

Scientists in drug discovery and development: We recognize scientists with drug discovery and development knowledge but limited computer programming skills as one type of users. We assume they will more likely access our system through a graphic user interface (GUI) such as a web browser. Figure 12 illustrates the interaction between

a drug discovery and development scientist and the system. While the user interacts with the system, he or she will provide system a text file containing one or more user posts along with a drug name. The GUI will map the drug name to a drug ID if it exists in our database and then pass the social media text and drug ID to the controller. The controller then passes the input from GUI to the model layer. After the processing with NLP and machine learning functions is complete, the controller generates a HTML page to visualize the NLP and serendipity prediction results in the GUI.

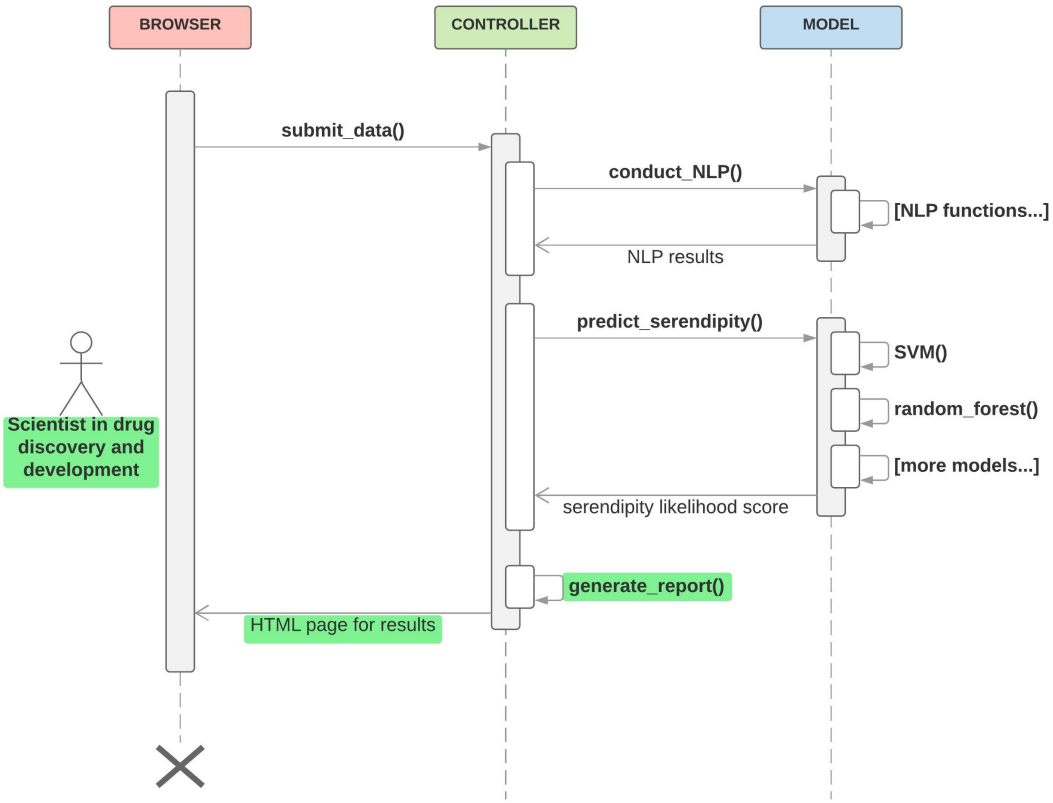

Figure 12: Interaction between a scientist in drug discovery and development and the system

Software developers: The second type of users are software developers who have computer programming skills but limited knowledge and experience with NLP and

machine learning. For these users, we provide a RESTful web interface that takes social media posts and drug ID from the user as inputs. The controller passes the input to the model layer for NLP and machine learning prediction and returns results in the JavaScript Object Notation (JSON) format, which is structured and commonly adopted for exchanging data between software applications. The overview of interaction is illustrated in Figure 13.

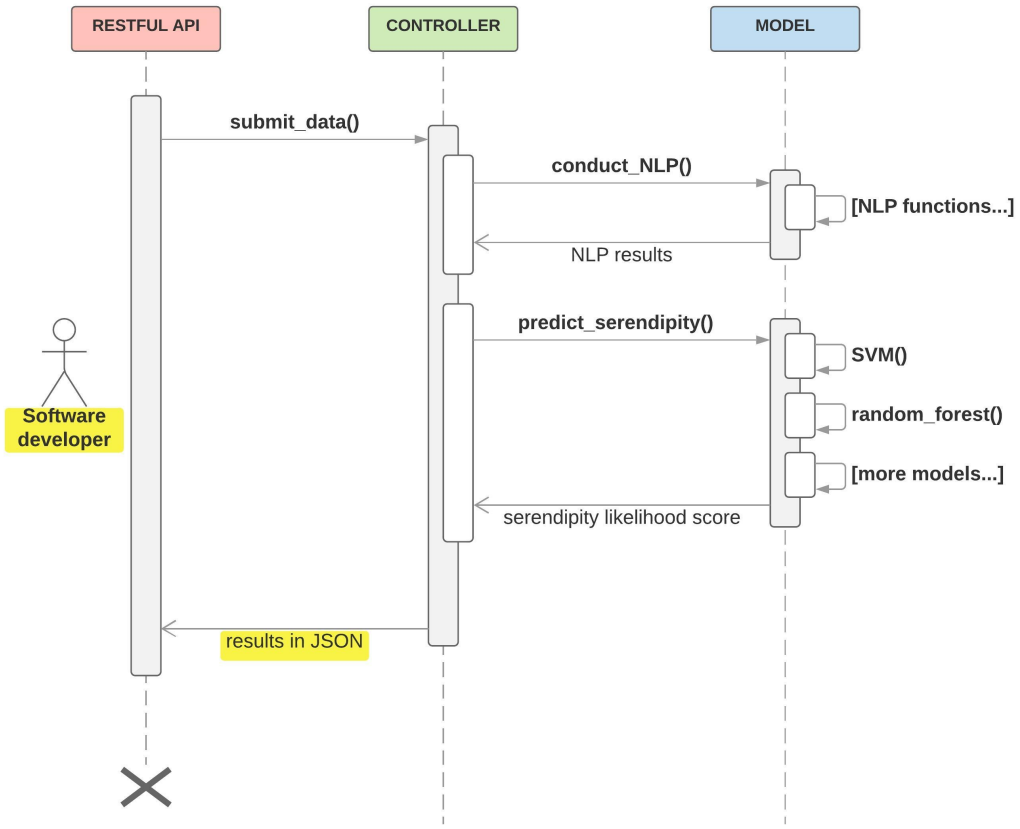

Figure 13: Interaction between software developer user and the system

5.4 Architecture and system components

We designed our system based on the MVC pattern [106] and client-server architecture [109]. The major components of the system include GUI, RESTful web

interface, NLP module, machine learning module, and controller. Several functions of the system depend on external APIs, libraries, and tools, including Stanford CoreNLP [32], MetaMap [33], Scikit-learn [104], and Keras [91]. The overview of the architecture is illustrated in Figure 14.

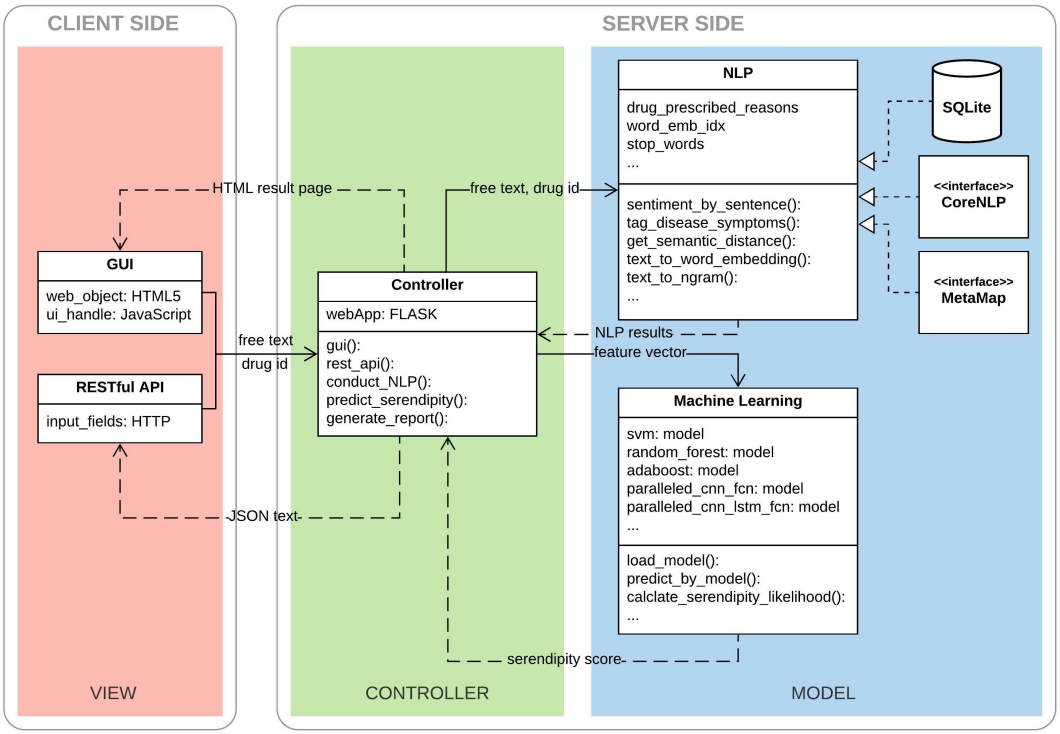

Figure 14: Overview of Serendipity system

GUI and RESTful web interface: These two user interfaces collect social media posts and the drug ID from the user. They are client side components of the system. The GUI askes the user to select the drug from a dropdown list and convert user choice to drug ID in our database at backend. Alternatively, the user can enter part of the drug's name and the GUI will interactively narrow down the list. For RESTful web interface, it receives drug ID directly from the user along with social media posts through HTTP

request. After receiving analytics results from the report, the GUI visualizes results as a HTML page and the RESTful web interface returns results in the JSON format.

Controller: The controller coordinates between user interfaces and analytics modules. As a server side component, it renders data between client-side components and analytics models. After completing the request from GUI or RESTful web interface, it will continue running as a background service to wait for the next request.

Model - NLP module: After receiving social media posts and the drug ID from the controller, the NLP module passes the social media text to the *Sentiment_by_sentence* function, which connects Stanford CoreNLP server to split social media text to sentences and calculate the sentiment score for each sentence. Then, sentences are passed to *Tag_disease_symptoms* function, which connects MetaMap server to map diseases and symptoms mentioned in each sentence to UMLS concepts. In the next, UMLS concepts associated with known indications of the drug are retrieved from the SQLite database [110], and *get_semantic_distance* function calculates the semantic distance between UMLS concepts found by MetaMap and these associated with the drug's known indications. Besides extracting semantic information, another task of NLP module is to generate features for each sentence. This is completed by *text_to_word_embedding* and *text_to_ngram* functions. In the end, sentiment score, UMLS concepts and associated information from MetaMap, drug's therapeutic areas (Section 3.2.5), and semantic distances are combined as non-text features for machine learning analysis. They were sent back to controller along with word embedding and n-gram features.

Model - Machine learning module: This module initializes five pre-trained prediction models, namely SVM, random forest, AdaBoost.M1, Paralleled CNN and

FCN, and Paralleled CNN-LSTM and FCN. These modules were built and trained using Scikit-learn [104] and Keras 2.0.8 [91] Python machine learning libraries. After receiving feature vector for a sentence from the Controller, each of the five models predicts the probability of mentioning serendipitous usages. The module calculates the average of the probabilities from five models and returns it to controller as the serendipity likelihood score.

5.5 Implementation

We implemented *Serendipity* in Python 3.6, Flask web development framework, and supporting libraries such as Scikit-learn and Keras. We also included MetaMap Web API and Stanford CoreNLP program in the software package.

We built the GUI in HTML5, which is a markup language to create web pages compatible with most recent computer operating systems and web browsers [111]. The GUI contains two web pages. The web_gui.html page (Figure 15 top) provides a text box for user to paste drug reviews from social media, an input box for users to enter the ID or name of the drug associated with the reviews. The input box also embeds a drop-down list which automatically suggests options when the user types drug name or ID. At the bottom of the page are a ‘Submit’ button to submit input data for analysis and a ‘Reset’ button to clear all current inputs so the user could enter new inputs. The report.html page (Figure 15 bottom) displays each sentence from user as a paragraph in ‘Social media comments’ window and highlights disease and symptoms that are different from drug’s

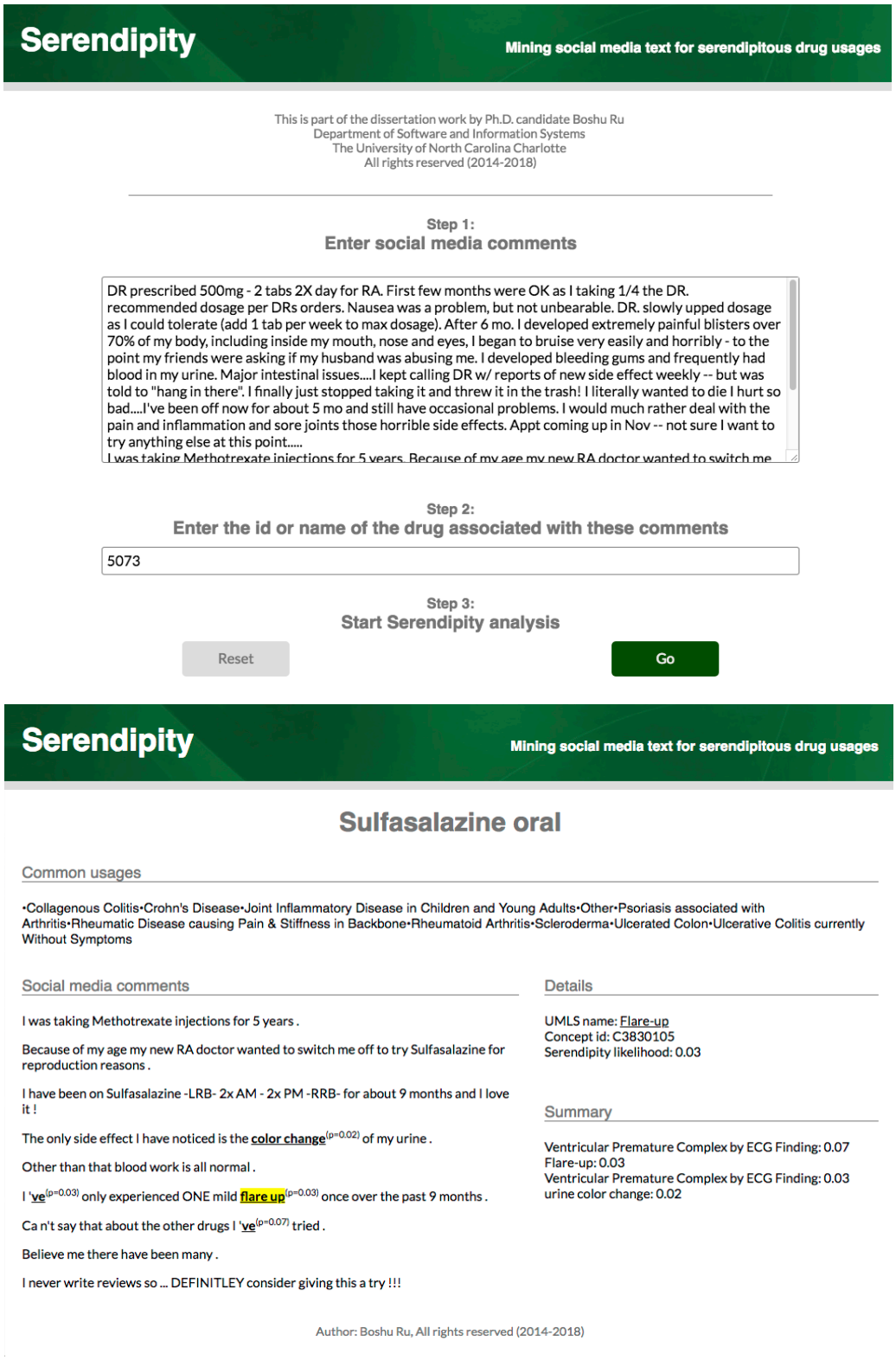

Figure 15: Graphic user interface

```
HTTP request:
localhost:5000/serendipity/api?d_id=5073&u_text=In%20addition%2
0to%20RA%20I%20have%20a%20history%20of%20IBS%20,%20sensitive%20
stomach%20and%20I%20have%20tolerated%20this%20medication%20well
%20.
-------------------------------------------------------------
{
  "common_use": "Rheumatoid Arthritis,...,Scleroderma",
  "drug_name": "Sulfasalazine oral",
  "sentence": [
    {
      "sentiment": 0.5,
      "text": "In addition to RA I have a history of IBS ,
    sensitive stomach and I have tolerated this medication
    well .",
      "concept": [
        {
          "cui": "C0022104",
          "distance": -0.6,
          "location": "38/3",
          "min distance": true,
          "negation": "0",
          "pos": "noun",
          "preferred term": "Irritable Bowel Syndrome",
          "semantic type": "dsyn",
          "trigger": "IBS"
        },
        {
          "cui": "C2004062",
          "distance": 1.0,
          "location": "27/7",
          "min distance": false,
          "negation": "0",
          "pos": "noun",
          "preferred term": "History of previous events",
          "semantic type": "fndg",
          "trigger": "history"
        }
      ],
      "prediction": [
        {
          "average": 0.06308634944375545,
          "model_ada": 0.18457321392843803,
          "model_rf": 0.0,
          "model_svm": 0.004685834402828326
        }
      ]
    }
```


Figure 16: Example of input and output in RESTful interface

Known indications in the bold font and displays the serendipity likelihood score predicted by models in the superscript. When the user clicks a highlighted word, the 'Details' window on the mid-right side of the page displays the UMLS mapping information and serendipity likelihood score for the disease or symptom. The 'Summary' window on the lower-right side lists all highlighted diseases and symptoms in the decreasing order of serendipity likelihood score. When the user clicked a highlighted item in 'Summary' window, the 'Social media comments' window will focus and highlight the item in the social media text. The report page also shows known indications of the drug, so the user could verify the predictions by their own pharmacy knowledge.

For RESTful interface, we built a Python program that listens HTTP request for two input arguments – social media text (u_text) and drug ID (d_id). The program will return analytics results in the JSON format (Figure 16).

We implemented the controller and the machine learning module with Python and its libraries. The controller utilized the Flask 0.12.2 framework to handle web service requests and responses. For the NLP module, *get_semantic_distance*, *text_to_word_embedding*, and *text_to_ngram* functions were developed in Python. The *Sentiment_by_sentence* function connects to the web interface of Stanford CoreNLP Java software that is running on the local computer. The *Tag_disease_symptoms* function calls a Java program to access the MetaMap service provided by the National Library of Medicine (NLM). We chose MetaMap service from the NLM because the terminology mapping tool requires 16 GB storage space to install on local machine, which occupies too much computing resource on the end user's computer. However, to use MetaMap service from NLM, the user needs a NLM account and have a stable internet connection.

5.6 Discussion

We implemented an open source application for mining serendipitous drug usages in social media. The application utilized natural language processing and machine learning methods explored in this dissertation work. The efforts we have taken are just the beginning of developing a software. The next step is to more specifically identify users and generate business or real-world use cases – for both user studies and potential customer interviews should proceed. We cautiously hypothesize two potential real world uses, without further validation.

A smart dashboard for clinical experts to annotate social media reflections. The mentions of serendipitous drug usages on social media can be potential clues to generate or validate drug repositioning hypotheses. Yet, such mentions need to be verified by clinical experts to exclude cases where the patient inaccurately described the medical events happened to him or her. Our analytics system and GUI could prioritize potential serendipitous usage mentions for clinical experts to verify, so that they could more effectively process patient reflections collected from social media.

An API for patient health forums to annotate user posts. After a user submitted a drug review post, the patient forums could pass the comments to Serendipity RESTful API. If the serendipitous usage likelihood score is higher than a threshold, the patient forum may pop additional questions for the user to verify if he or she indicated the drug also improves his or her comorbid condition. With only a few more seconds spent by each user, the patient forum could effectively collect annotated data for serendipitous drug usages.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

In the past decade, a critical mass of patient-reported medication outcomes data has been accumulated on social media. Many discussions in social media sites mention serendipitous drug usages, which could be useful hints for drug repositioning researchers – in the sense of complementing other data sources to generate or validate repositioning hypotheses. Comparing to traditional data sources such as EHR, claims, FAERS, and surveys, social media data contain a large volume of information voluntarily contributed by patients not limited to one geographic location or healthcare provider. Yet, the colloquial language makes social media data difficult for computers to understand the semantic contents. Moreover, inaccuracy in self-reported outcomes and the missing of contextual information such as co-prescribed drugs introduce noises.

This dissertation responded to these challenges by exploring natural language processing (NLP) and machine learning methods to detect discussions of serendipitous drug usages in social media posts, which is otherwise difficult to collect and analyze.

We started with a content analysis on the discussions of four diseases and 11 drugs on WebMD, PatientsLikeMe, YouTube (video comments), and Twitter [16]. We found patient health forums like WebMD the best social media data source for patient-reported medication outcomes in terms of data quality and we manually identified several drug reviews mentioning serendipitous usages.

Then, we explored NLP and machine learning methods to identify serendipitous drug usages from patient health forum. We designed information filters that leverages biomedical named entity recognition and normalization tool (MetaMap [33]), sentiment

analyzer (Stanford CoreNLP [32]), and medical ontology (SNOMED CT [34]) to extract potential serendipitous drug usages from social media text. We curated the first gold standard dataset for predicting serendipitous drug usages, which consists 447 sentences from WebMD that mentioned serendipitous drug usages and 15,267 sentences that did not. Then, we applied n-grams and machine learning algorithms, namely SVM, random forest, and AdaBoost.M1, as well as medical knowledge in the feature construction and modeling process. Our best model had AUC=0.937, Precision=0.811, Recall=0.476. Several predictions, including metformin and bupropion for obesity, tramadol for depression, and ondansetron for irritable bowel syndrome with diarrhea, are also supported by recent biomedical research publications.

Afterward, we explored deep learning models for the same prediction task. We constructed four deep learning models, using three types of neural networks – Convolutional Neural Network (CNN), Fully-connected Neural Network (FCN), and Long Short Term Memory network (LSTM). We examined model configuration, hyper parameters, prediction power, and complexity. The results show adding context information such as drug therapeutic areas to machine learning features is helpful to prevent models from overfitting. But deep learning models may not outperform traditional models in the presence of extremely imbalanced data.

In the end, we implemented an open source application for scientists in drug discovery and development and software developers to utilize most of this dissertation work without advanced NLP and machine learning skills. The application takes social media posts and the drug name as inputs and returns NLP and machine learning prediction results either in an interactive report page, or in a JSON format file. By

leveraging web technology of HTML5 and RESTful API, the application minimizes the need of configuration and computation power from the user.

This dissertation represents only our initial exploration of mining patient-reported medication outcomes in social media to identify serendipitous drug usages. Our work has several limitations that leave possibilities for further research, as discussed in the following:

We used MetaMap to identify disease and symptoms in social media text and map them to standard medical terminology. However, MetaMap and many other tools currently available to use, are designed either for clinical text or scientific literature, whose writing style is formal and the description of outcomes is more accurate than patient-reported medication outcomes in the social media. In our empirical study (Chapter 3), we found the current tools can be the performance limiting step due to the lack of customization for social media data. We believe more powerful named entity recognition and normalization tools specialized in social media text could be expected to improve the performance of the current system and is under development in the lab.

Construction of machine learning features from social media text also requires further efforts. We explored commonly used text mining methods, such as n-grams and word embedding in the current research. More sophisticated text features, such as part-of-speech, sentence syntactic structure (shallow parsing tree), and semantic topics retrieved from topic modeling might worth further investigation [112-114].

Lack of true serendipitous drug usage cases in the gold standard data remains an obstacle for improving the performance of machine learning models, despite efforts we took such as data sampling and cost sensitive learning methods. We will identify more

positive serendipitous drug usage cases from social media and re-evaluate the performance of our models.

Last but not least, the software application we developed is only a prototype. We have not conducted user review and test to identify more specific user group, generate business use cases, and evaluate the utility of this tool to scientists in drug discovery and development and software developers. In addition, the platform, architecture, distribution methods are subject to further investigation in terms of security and scalability. These steps are necessary to complete the software development cycle before we can release it.

REFERENCES

[1] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature Review Drug Discovery,* vol. 3, pp. 673-683, 2004.

[2] J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug–disease relationships for computational drug repositioning," *Briefings in Bioinformatics,* vol. 12, pp. 303-311, 2011.

[3] L. Yao, Y. Zhang, Y. Li, P. Sanseau, and P. Agarwal, "Electronic health records: Implications for drug discovery," *Drug Discovery Today,* vol. 16, pp. 594-599, 2011.

[4] K. L. Shandrow. (2016, 02/22/2016). *The Hard Truth: What Viagra Was Really Intended For*. Available: http://www.entrepreneur.com/article/254908

[5] C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis, "Literature mining, ontologies and information visualization for drug repurposing," *Briefings in Bioinformatics,* vol. 12, pp. 357-368, 2011.

[6] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. Abbas, S. J. Hufeisen*, et al.*, "Predicting new molecular targets for known drugs," *Nature,* vol. 462, pp. 175-181, 2009.

[7] P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon*, et al.*, "Use of genome-wide association studies for drug repositioning," *Nature Biotechnology,* vol. 30, pp. 317-320, 2012.

[8] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PLoS ONE,* vol. 4, p. e6536, 2009.

[9] U.S. National Library of Medicine. (2016, 09/29/2016). *MEDLINE Fact Sheet*. Available: https://www.nlm.nih.gov/pubs/factsheets/medline.html

[10] J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner, "Knowledge discovery by automated identification and ranking of implicit relationships," *Bioinformatics,* vol. 20, pp. 389-398, 2004.

[11] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology,* vol. 7, p. 496, 2011.

[12] J. S. Shim and J. O. Liu, "Recent advances in drug repositioning for the discovery of new anticancer drugs," *International Journal of Biological Sciences,* vol. 10, pp. 654-63, 2014.

[13] P. Khatri, S. Roedder, N. Kimura, K. De Vusser, A. A. Morgan, Y. Gong, *et al.*, "A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation," *The Journal of Experimental Medicine,* vol. 210, pp. 2205-2221, 2013.

[14] H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, *et al.*, "Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality," *Journal of the American Medical Informatics Association,* vol. 22, pp. 179-191, 2014.

[15] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics,* vol. 13, pp. 395-405, 2012.

[16] B. Ru, K. Harris, and L. Yao, "A Content Analysis of Patient-Reported Medication Outcomes on Social Media", in *Proceedings of IEEE 15th International Conference on Data Mining Workshops*, Atlantic City, NJ, USA, 2015, pp. 472-479.

[17] Q. T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *Journal of the American Medical Informatics Association,* vol. 13, pp. 24-29, 2006.

[18] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, "Social media mining for drug safety signal detection", in *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, Maui, HI, USA, 2012, pp. 33-40.

[19] M. Hahsler. (2015, 02/13/2015). *Probabilistic Comparison of Commonly Used Interest Measures for Association Rules.* Available: http://michael.hahsler.net/research/association_rules/measures.html

[20] A. Yates and N. Goharian, "ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites," *Advances in Information Retrieval,* pp. 816-819, 2013.

[21] M. Z. Knezevic, I. C. Bivolarevic, T. S. Peric, and S. M. Jankovic, "Using Facebook to increase spontaneous reporting of adverse drug reactions," *Drug Safety,* vol. 34, pp. 351-352, 2011.

[22] G. E. Powell, H. A. Seifert, T. Reblin, P. J. Burstein, J. Blowers, J. A. Menius*, et al.*, "Social media listening for routine post-marketing safety surveillance," *Drug Safety,* vol. 39, pp. 443-454, 2016.

[23] B. Honigman. (2013, 06/17/2015). *24 Outstanding Statistics & Figures on How Social Media has Impacted the Health Care Industry*. Available: https://getreferralmd.com/2013/09/healthcare-social-media-statistics/

[24] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng*, et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, 2013, pp. 1631-1642.

[25] F. Ochoa-Cortes, A. Liñán-Rico, K. A. Jacobson, and F. L. Christofi, "Potential for Developing Purinergic Drugs for Gastrointestinal Diseases," *Inflammatory bowel diseases,* vol. 20, pp. 1259-1287, 2014.

[26] Centers for Disease Control and Prevention. (2014, 02/04/2015). *Diabetes Latest*. Available: http://www.cdc.gov/features/diabetesfactsheet/

[27] D. E. Kanouse and L. R. Hanson Jr, "Negativity in evaluations," in *Attribution: Perceiving the causes of behavior*, 1 ed Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1987, pp. 47-62.

[28] A. K. Lalwani, *Negativity and Positivity Biases in Product Evaluations: The Impact of Consumer Goals and Prior Attitudes*: ProQuest, 2006.

[29] H. W. Walling and B. L. Swick, "Update on the management of chronic eczema: new approaches and emerging treatment options," *Clinical, cosmetic and investigational dermatology : CCID,* vol. 3, pp. 99-117, 07/28/2010 2010.

[30] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, "Exploiting emoticons in sentiment analysis", in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 703-710.

[31] O. Tsur, D. Davidov, and A. Rappoport, "ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews", in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington, DC, USA, 2010, pp. 162-169.

[32] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit", in *The 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, MD, USA, 2014, pp. 55-60.

[33] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association,* vol. 17, pp. 229-236, 2010.

[34] U.S. National Library of Medicine. (2016, 08/03/2015). *SNOMED CT*. Available: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[35] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of Biomedical Informatics,* vol. 40, pp. 288-299, 2007.

[36] N. H. Shah and M. A. Musen, "UMLS-Query: a perl module for querying the UMLS", in *AMIA Annual Symposium*, Washington, DC, USA, 2008, pp. 652-656.

[37] L. Yao and A. Rzhetsky, "Quantitative systems-level determinants of human genes targeted by successful drugs," *Genome Research,* vol. 18, pp. 206-213, 2008.

[38] S. Frantz, "Drug discovery: playing dirty," *Nature,* vol. 437, pp. 942-943, 2005.

[39] L. Pleyer and R. Greil, "Digging deep into “dirty” drugs–modulation of the methylation machinery," *Drug Metabolism Reviews,* vol. 47, pp. 252-279, 2015.

[40] J. Fürnkranz, "A study using n-gram features for text categorization," *Austrian Research Institute for Artifical Intelligence,* vol. 3, pp. 1-10, 1998.

[41] I. Feinerer and K. Hornik, "tm: text mining package," *R package version 0.5-7.1,* 2012.

[42] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*: Springer Science & Business Media, 2013.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.

[44] S. Ali and K. A. Smith-Miles, "Improved support vector machine generalization using normalized input space", in *Proceedings of the 19th Australasian Joint Conference on Artificial Intelligence*, Hobart, Australia, 2006, pp. 362-371.

[45] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[46] L. Breiman, "Random forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[47] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm", in *Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy, 1996, pp. 148-156.

[48] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, p. 27, 2011.

[49] J. R. Quinlan, *C4.5: programs for machine learning*: Elsevier, 2014.

[50] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, pp. 1263-1284, 2009.

[51] R. Batuwita and V. Palade, "Efficient resampling methods for training support vector machines with imbalanced datasets", in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Barcelona, Spain, 2010, pp. 1-8.

[52] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering,* vol. 30, pp. 25-36, 2006.

[53] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria", in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, 2004, pp. 69-78.

[54] L. I. Igel, A. Sinha, K. H. Saunders, C. M. Apovian, D. Vojta, and L. J. Aronne, "Metformin: an old therapy that deserves a new indication for the treatment of obesity," *Current Atherosclerosis Reports,* vol. 18, pp. 1-8, 2016.

[55] A. R. Desilets, S. Dhakal-Karki, and K. C. Dunican, "Role of metformin for weight management in patients without type 2 diabetes," *Annals of Pharmacotherapy,* vol. 42, pp. 817-826, 2008.

[56] G. Paolisso, L. Amato, R. Eccellente, A. Gambardella, M. R. Tagliamonte, G. Varricchio*, et al.*, "Effect of metformin on food intake in obese subjects," *European Journal of Clinical Investigation,* vol. 28, pp. 441-446, 1998.

[57] L. Peirson, J. Douketis, D. Ciliska, D. Fitzpatrick-Lewis, M. U. Ali, and P. Raina, "Treatment for overweight and obesity in adult populations: a systematic review and meta-analysis," *CMAJ Open,* vol. 2, pp. E306-E317, 2014.

[58] M. S. McDonagh, S. Selph, A. Ozpinar, and C. Foley, "Systematic review of the benefits and risks of metformin in treating obesity in children aged 18 years and younger," *JAMA Pediatrics,* vol. 168, pp. 178-184, 2014.

[59] P. L. Tenore, "Psychotherapeutic benefits of opioid agonist therapy," *Journal of Addictive Diseases,* vol. 27, pp. 49-65, 2008.

[60] T. Tetsunaga, T. Tetsunaga, M. Tanaka, and T. Ozaki, "Efficacy of tramadol–acetaminophen tablets in low back pain patients with depression," *Journal of Orthopaedic Science,* vol. 20, pp. 281-286, 2015.

[61] A. L. Stoll and S. Rueter, "Treatment augmentation with opiates in severe and refractory major depression," *American Journal of Psychiatry,* vol. 156, p. 2017, 1999.

[62] F. L. Greenway, K. Fujioka, R. A. Plodkowski, S. Mudaliar, M. Guttadauria, J. Erickson*, et al.*, "Effect of naltrexone plus bupropion on weight loss in overweight and obese adults (COR-I): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial," *The Lancet,* vol. 376, pp. 595-605, 2010.

[63] K. M. Gadde, C. B. Parker, L. G. Maner, H. R. Wagner, E. J. Logue, M. K. Drezner*, et al.*, "Bupropion for weight loss: an investigation of efficacy and tolerability in overweight and obese women," *Obesity Research,* vol. 9, pp. 544-551, 2001.

[64] J. W. Anderson, F. L. Greenway, K. Fujioka, K. M. Gadde, J. McKenney, and P. M. O'Neil, "Bupropion SR Enhances Weight Loss: A 48-Week Double-Blind, Placebo-Controlled Trial," *Obesity Research,* vol. 10, pp. 633-641, 2002.

[65] A. K. Jain, R. A. Kaplan, K. M. Gadde, T. A. Wadden, D. B. Allison, E. R. Brewer*, et al.*, "Bupropion SR vs. placebo for weight loss in obese patients with depressive symptoms," *Obesity Research,* vol. 10, pp. 1049-1056, 2002.

[66] K. Garsed, J. Chernova, M. Hastings, C. Lam, L. Marciani, G. Singh*, et al.*, "A randomised trial of ondansetron for the treatment of irritable bowel syndrome with diarrhoea," *Gut,* vol. 63, pp. 1617-1625, 2014.

[67] R. Leaman, R. I. Doğan, and Z. Lu, "DNorm: disease name normalization with pairwise learning to rank," *Bioinformatics,* vol. 29, pp. 2909-2917, 2013.

[68] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[69] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification", in *AAAI*, 2015, pp. 2267-2273.

[70] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A C-LSTM neural network for text classification," *arXiv preprint arXiv:1511.08630,* 2015.

[71] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning", presented at the Proceedings of the 25th international conference on Machine learning, Helsinki, Finland, 2008.

[72] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", in *Advances in neural information processing systems*, 2013, pp. 3111-3119.

[73] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[74] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification", in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1555-1565.

[75] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[76] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks", in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 959-962.

[77] A. Chapman. (2015, 04/05/2018). *Bag of Words Meets Bags of Popcorn - Use Google's Word2Vec for movie reviews*. Available: https://www.kaggle.com/c/word2vec-nlp-tutorial

[78] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora", in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[79] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *J Am Med Inform Assoc,* vol. 22, pp. 671-81, 2015.

[80] Y. Kim, "Convolutional Neural Networks for Sentence Classification", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746-1751.

[81] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep Learning for Extreme Multi-label Text Classification", in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 115-124.

[82] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification", in *Advances in neural information processing systems*, 2015, pp. 649-657.

[83] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[84] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions", 2015.

[85] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal*, et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316,* 2016.

[86] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model", in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[87] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks,* vol. 18, pp. 602-610, 2005.

[88] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[89] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, pp. 1735-1780, 1997.

[90] H. Li, M. R. Min, Y. Ge, and A. Kadav, "A Context-aware Attention Network for Interactive Question Answering", in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 927-935.

[91] (2017, 04/05/2018). *Keras: The Python Deep Learning library*. Available: https://keras.io/

[92] (2017). *Computation using data flow graphs for scalable machine learning*. Available: https://github.com/tensorflow/tensorflow

[93] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines", in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807-814.

[94] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249-256.

[95] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.

[96] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research,* vol. 134, pp. 19-67, 2005.

[97] M. Kuhn and K. Johnson, *Applied predictive modeling* vol. 26: Springer, 2013.

[98] C. Cortes, M. Mohri, and A. Rostamizadeh, "L 2 regularization for learning kernels", in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 109-116.

[99] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026-1034.

[100] M. Pumperla. (2017, 04/05/2018). *Keras + Hyperopt: A very simple wrapper for convenient hyperparameter optimization*. Available: https://github.com/maxpumperla/hyperas

[101] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms", in *Proceedings of the 12th Python in Science Conference*, 2013, pp. 13-20.

[102] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization", in *Advances in neural information processing systems*, 2011, pp. 2546-2554.

[103] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves", in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233-240.

[104] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel*, et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research,* vol. 12, pp. 2825-2830, 2011.

[105] Treato.com. (2018, 04/05/2018). *Technology - Understanding patient language*. Available: https://corp.treato.com/technology

[106] G. E. Krasner and S. T. Pope, "A description of the model-view-controller user interface paradigm in the smalltalk-80 system," *Journal of object oriented programming,* vol. 1, pp. 26-49, 1988.

[107] A. Ronacher, G. Brandl, A. Zapletal, A. Afshar, C. Edgemon, C. Grindstaff*, et al.*, "Flask (a Python microframework)," http://flask/. *pocoo. org. Acessado em,* vol. 6, p. 2014, 2010.

[108] L. Richardson and S. Ruby, *RESTful web services*: " O'Reilly Media, Inc.", 2008.

[109] A. Berson, *Client-server architecture*: McGraw-Hill, 1992.

[110] M. Owens and G. Allen, *SQLite*: Springer, 2010.

[111] B. Frain, *Responsive web design with HTML5 and CSS3*: Packt Publishing Ltd, 2012.

[112] X. Liu and H. Chen, "A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports," *J Biomed Inform,* vol. 58, pp. 268-79, 2015.

[113] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *J Biomed Inform,* vol. 53, pp. 196-207, 2015.

[114] M. Yang, M. Kiang, and W. Shang, "Filtering big data from social media--Building an early warning system for adverse drug reactions," *J Biomed Inform,* vol. 54, pp. 230-40, 2015.