# Predicting Drug-Drug Interactions from Text using NLP and Privileged Information

Alexander L. Hayes
Indiana University
hayesall@indiana.edu

Savannah Smith
Valparaiso University
savannah.smith@valpo.edu

Devendra Singh Dhami
Indiana University
ddhami@indiana.edu

Sriraam Natarajan
Indiana University
natarasr@indiana.edu

## ABSTRACT

*Objective*: The medication prescribed to help treat a condition can also be detrimental to a person's health: drug-drug interactions or adverse drug events (ADEs) can be deadly, and after four concurrent medications the chance of encountering one increases exponentially. As further motivation, it is nearly impossible for every drug-drug combination to be represented during clinical trials, nor is it possible to encapsulate the breadth of genetic diversity between people that will inevitibably use these drugs. Among the possible solutions to this problem, we applied machine learning and natural language processing (NLP) to data from drug labels and scientific studies. Furthermore, we show that privileged information from medical blogs can be incorporated to improve our model. Using this method, we can accurately predict drugs and their events, and if used effectively, this can improve safety and allow patients to make informed health decisions.

*Methods*: We mined data from three sources: the openFDA drug label database, PubMed, and a variety of medical blogs (Dailystrength, et al.). After building our dataset, we labeled for keywords that could be processed with Stanford NLP (”interacts with”, ”inhibits cytochrome p450 production”), then applied a Recurrent Neural Network (RNN) to model the ADEs.

*Results*: We got results and they were pretty cool.

*Conclusions*: We drew a conclusion from our pretty cool results.

## 1. INTRODUCTION

Adverse drug-drug interactions have become an increasing problem under polypharmacy (typically described as being prescribed more than four medications simultaneously), and is a great concern for people aged 65 and older that commonly find themselves in this situation. Machine learning has already been applied to the medical field in multiple ways, including protein shape prediction, adverse event prediction, and risk identification. We use similar techniques to predict adverse events from drug-drug interactions using Natural Language Processing (NLP), PubMed Abstract data, the openFDA drug labeling database, and information scraped from websites and blogs.

Many researchers have tried to find ways to minimize or even eliminate ADEs. This is difficult especially in the sense that each time a new drug is created it has a chance of causing a bad reaction; rather the reaction is due to that single drug or being mixed with another drug. This paper focuses on the extraction of drug-drug interactions and their events.

Despite the many researchers working to solve the ADEs issue, ADEs continue to occur, day in and day out. It is difficult to find the causation relationship of drug-drug interactions and their events with confidence.By confidence, I mean having used multiple data sets, data bases, and social blogs in order to have more support to uphold our final model.In this paper, they extracted data from numerous professional databases and social blogs, to either confirm or deny each other, in order to more confidently support their findings. Current researchers investigate single ADEs where as the author of this paper expanded the research on single ADEs to find combinations of drugs (drug-drug) interactions and their events. The author researched ADEs in order to address and inform people on this major concern that accounts for many injuries and deaths each year. Finally, The contribution of this paper is to use machine learning and NLP to extract text from the web. The primary contribution of this work is to expand on a probabilistic method for summarizing what the (research) community knows about drug-drug interactions and their events.

## 2. RELATED WORK

Early studies on Adverse Drug Events (ADEs) [1, 2, 3] found how dangerous of a problem they were and had potential to be. Evidence suggests there have been over two million ADEs yearly with death tolls reaching between 44,000 and 98,000 lives. Since some of the original research, there have been many proposed solutions for how to mediate these risks: some at the drug creation level [4, 5] and some after drugs are on the market. Sometimes it's only possible to find ADEs when the drugs are ”out in the wild” due to the difficulty of detection during clinical trials. Because of the space between people who may experience a reaction, this

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2016 ACM. ISBN 123-4567-24-567/08/06.
DOI: 10.475/123_4

becomes an apt for machine learning. Previous work has focused on adverse reactions using Natural Language Processing (NLP) [6, 7].

## 2.1 Process of Extracting ADEs

This section focuses on using Machine learning to identify and extract possible ADEs from text.
"Efficient strategies for identification and extraction of information about potential adverse drug events from free-text resources are needed to support pharmacovigilance research and pharmaceutical decision making" [8]. This paper focuses on using Machine Learning to extract possible ADEs from MEDLINE case reports. We are also using Machine learning to extract possible ADEs but we are extracting from medical abstracts and social blogs and plan on later comparing them to either confirm or deny drug-drug interactions and their events. "The contribution of this paper is to propose a new method that integrates Natural Language Processing (NLP) and Machine Learning (ML) to extract adverse drug events from published text" [9]. This paper was written by professor Natarajan himself along with a few other researchers. It gives a detail description of their purpose, procedures, and conclusion in relation with extracting ADEs from text. Natarajan's paper is detailing ADE pairs where as we our detailing drug-drug interactions and their events.

## 2.2 Background Knowledge

This section contains basic skills or knowledge you need in order to Extract ADEs.For example, before diving into creating the concluding model some knowledge on programming, essentials of databases, and the use and understanding of algorithms would be helpful.
Induction logic programming is a subfield of machine learning which uses logic programming as a uniform representation for examples, background knowledge and hypothesis [10]."Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look" [11]. In addition to these, you must familiarize yourself with things such as medical blogs, medical abstracts, PubMed, Statistical Relational Learning, FDA, and DailyMed.

AND Backward Citation Tree.PNG AND Backward Citation Tree.PNG

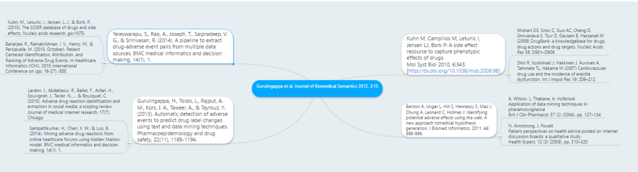
Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2015). The SIDER database of drugs and side effects. Nucleic acids research, gkv1075.
Banerjee, R., Ramakrishnan, I. V., Henry, M., & Perciavalle, M. (2015, October). Patient Centered Identification, Attribution, and Ranking of Adverse Drug Events. In Healthcare Informatics (ICHI), 2015 International Conference on (pp. 18-27). IEEE.
Yeleswarapu, S., Rao, A., Joseph, T., Saipradeep, V. G., & Srinivasan, R. (2014). A pipeline to extract drug-adverse event pairs from multiple data sources. BMC medical informatics and decision making, 14(1), 1.
Lardon, J., Abdellaoui, R., Bellet, F., Asfari, H., Souvignet, J., Texier, N., ... & Bousquet, C. (2015). Adverse drug reaction identification and extraction in social media: a scoping review. Journal of medical Internet research, 17(7). Chicago
Sampathkumar, H., Chen, X. W., & Luo, B. (2014). Mining adverse drug reactions from online healthcare forums using hidden Markov model. BMC medical informatics and decision making, 14(1), 1.
Gurulingappa, H., Toldo, L., Rajput, A. M., Kors, J. A., Taweel, A., & Tayrouz, Y. (2013). Automatic detection of adverse events to predict drug label changes using text and data mining techniques. Pharmacoepidemiology and drug safety, 22(11), 1189-1194.
Gurulingappa et al. Journal of Biomedical Semantics 2012, 3:15
Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P: A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010, 6:343. [http://dx.doi.org/10.1038/msb.2009.98]
Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901–D906
Shiri R, Koskimaki J, Hakkinen J, Auvinen A, Tammela TL, Hakama M (2007) Cardiovascular drug use and the incidence of erectile dysfunction. Int J Impot Res 19: 208–212
Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, Leonard C, Holmes J: Identifying potential adverse effects using the web: A new approach tomedical hypothesis generation. J Biomed Informatics. 2011, 44: 989-996.
A. Wilson, L. Thabane, A. Holbrook Application of data mining techniques in pharamcovigilance Brit J Clin Pharmacol, 57 (2) (2004), pp. 127–134
N. Armstrong, J. Powell Patient perspectives on health advice posted on Internet discussion boards: a qualitative study Health Expect, 12 (3) (2009), pp. 313–320

**Figure 1: Forward and Backword Citation Tree**

# 3. METHODS

We decided on three main bodies of information to make predictions with: PubMed Abstracts, Label information from the FDA, and social/medical blogs from the perspectives of people taking these drugs.

However, before we could do any of these, we needed a list of drugs to study. This list of drugs was created using a bash script (builddruglist.sh) that pulled the information from rxlist.com. Because the information on rxlist.com is updated over time (between Thursday, June 9, 2016 and Tuesday, June 14, 2016: 10 new drugs were added) we needed to account for new additions.

In order to make corrections, another script (fixlist.sh) would download a fresh copy of the list, check it against the log file generated by fdainteractions.sh, and return a file that contained all drugs not present in the log file.

Another script used this list of drugs to query the openFDA drug label database (fdainteractions.sh). The script searched the database for generic drugs matching the input, if it was not a generic form there was a check to see if it was a brand name, then if it matched neither category it was dubbed "UNKNOWN" and grouped respectively. This was done as a way to minimize redundant information: pulling in information about the generic form of a drug also pulled in information about the brand name versions. For each generic name queried there were three categories we were interested in: drug_interactions, adverse_reactions, and warnings_and_cautions. Between these three categories we expected to get enough information to train the model to link adverse reactions with drug-drug reactions.

Meanwhile, a script called DailyFinal.py was written to automatically extract social blogs (discussions) of our choice. The social blog site used was called DailyStrength.org. In order to begin solving this problem efficiently one step at a time, the author decided to focus on a max amount drug combinations. A set list of drugs were considered in this research. A script extracting text from social blogs was created before the specific drugs was even considered. Then, After all the text was extracted into a file, there was a more in depth search of drug-drug combinations and their effects within the social blogs.
For this paper's purposes, the social blog aspect of the research will not be used at this moment. After the text was extracted from both the social blog(s) and databases, someone labeled each text as positive or negative. Positive being it is an ADE and negative describing that the text is not an ADE. Then the text was ran through the Natural Language Processor (NLP) to be analyzed. After this, using the labels, the NLP results, and a background file, the author was able to train a RDN Boosting algorithm. The plan was to use this trained boosting algorithm and test it with the PubMed data.

Now we describe the models that we used in this paper for the classification phase. The first model used is Long Short term memory(LSTM) [12]. These are a special kind of recurrent neural networks that can learn long-term dependencies. To understand long term dependencies consider the example of the human brain. When a human tries to solve a presented problem, instead of solving it from scratch, we tend to think about similar problems we might have solved in the past and try to figure out a problem iteratively from the previous solution if it exists. Thus we are dependent on the solution stored in our memory. Also, there is no limit to the time frame of the solved problem i.e. it can be way in the past. Thus the term 'long term dependency'. LSTM overcomes this long term dependency problem with a chain of repeating modules. This is shown in figures 2 and 3 adapted from [13].

As we can see that the middle module, also known as the repeating module, has a simple 1 layer structure in case

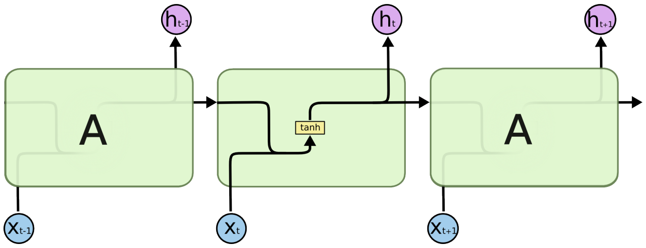

**Figure 2: Standard RNN**

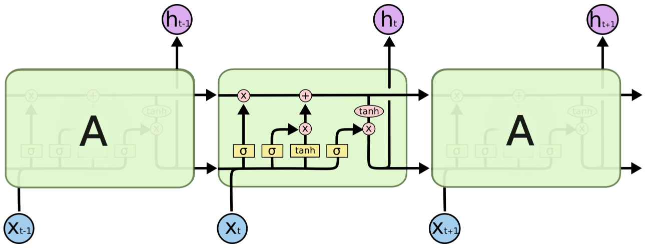

**Figure 3: LSTM**

of RNN but a complex 4 layer structure in case of LSTM thereby helping the network to remember better.

The second model we discuss here is called RDN boost [14] which means boosting relational dependency networks. Dependency networks [15] learn conditional dependencies of variables independently and then approximate the joint distribution over the variables as a product of these conditional dependencies.Relational dependency networks extend the dependency networks to relational domains.

## 4. FINDINGS AND DISCUSSION

After gathering and computing all necessary input files, the recurrent neural network was ready to be ran. The files needed to get results from the RNN are the background file, training set, positive examples, and negative examples. There were 2,000 openFDA files extracted and implemented, 250 positively labeled drug-drug interactions, and approximately 100 negatively labeled drug-drug interactions. The previously retrieved openFDA data was ran through a natural language processor in order to obtain the training set for the LSTM network. In this paper, tHe openFDA data was divided into 3 parts,the training set (50%), validation set (10%), and the test set (40%). After the algorithm was complete are results were as follows:

- Training Accuracy: 84.6%
- Validation Accuracy: 100% (its a small validation set)
- Test Accuracy: 73.1% (expected for such a less amount of data)

There are several possible directions for future work. In this work, we analyzed drug-drug interactions using RNN but the plan is to explore other models and compare the results to determine the most effective. Another possibility is to including more data sets when training the the model in order to have more efficient results.

## 5. REFERENCES

[1] Linda T Kohn, Janet M Corrigan, Molla S Donaldson, et al. *To err is human:: building a Safer Health System*, volume 6. National Academies Press, 2000.

[2] Jason Lazarou, Bruce H Pomeranz, and Paul N Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279(15):1200–1205, 1998.

[3] Jerry H Gurwitz, Terry S Field, Jerry Avorn, Danny McCormick, Shailavi Jain, Marie Eckler, Marcia Benser, Amy C Edmondson, and David W Bates. Incidence and preventability of adverse drug events in nursing homes. *The American journal of medicine*, 109(2):87–94, 2000.

[4] Liang-Chin Huang, Xiaogang Wu, and Jake Y Chen. Predicting adverse side effects of drugs. *BMC genomics*, 12(5):1, 2011.

[5] Marcel J De Groot. Designing better drugs: predicting cytochrome p450 metabolism. *Drug discovery today*, 11(13):601–606, 2006.

[6] Carol Friedman. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In *Artificial Intelligence in Medicine*, pages 1–5. Springer, 2009.

[7] Phillip Odom, Vishal Bangera, Tushar Khot, David Page, and Sriraam Natarajan. Extracting adverse drug events from text using human advice. In *Artificial Intelligence in Medicine*, pages 195–204. Springer, 2015.

[8] Harsha Gurulingappa. Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, pages 1–10, 2012.

[9] Tushar Khot Jose Picado Anurag Wazalwar Vitor Santos Costa David Page Sriraam Natarajan, Vishal Bangera and Michael Caldwell. A novel method for adverse drug event extraction from text, 2016.

[10] Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8:295–318, 1991.

[11] SAS Institute. Machine learning: What it is and why it matters.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] SAS Institute. Machine learning: What it is and why it matters.

[14] Sriraam Natarajan, Tushar Khot, Kristian Kersting, Bernd Gutmann, and Jude Shavlik. Boosting relational dependency networks. In *Online Proceedings of the International Conference on Inductive Logic Programming 2010*, pages 1–8, 2010.

[15] David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.