Dissertation Critique: Exploring Machine Learning Techniques Using Patient Interactions In Online Health Forums to Classify Drug Safety

Christopher Jeschke CJESCHK2@JHU.EDU
Engineering for Professionals
Johns Hopkins University
Elkridge, MD 20175, USA

Editor: n/a

Abstract

Patient generated health data represents an area of active research interest for its potential applications in improving the public health. The study of Pharmacovigilance is one such area, focused on monitoring drugs once they have been released to market. Dr. Brant Chee's 2011 dissertation applying machine learning techniques to patient messages in online health forums explores how watch list drugs from the United States Food and Drug Administration can be detected via these forum messages, ultimately with the intent to alert consumers to drug safety concerns.

Keywords: Drug Safety, Pharmacovigilance, NLP

1. Summary of Research

Dr. Brant Chee's 2011 dissertation *Exploring Machine Learning Techniques Using Patient Interactions in Online Health Forums to Classify Drug Safety* describes Chee's research in applying natural language processing (NLP) techniques in conjunction with Naive Bayes and Support Vector Machine classifiers to identify candidate *watch list* drugs from online patient forums. Watch list drugs are those drugs identified by the United States Food and Drug Administration (FDA) as presenting a significant health or safety risk to drug consumers, thereby prompting regulatory action to better inform the consumer or directly protect the consumer by removing the drug from market or reducing its accessibility. Chee's dissertation seeks to answer the specific questions:

- Can Machine Learning classification methods using text features extracted from online health forums be used to identify FDA watch list drugs?
- Is the sentiment of the forum message useful in identifying these drugs?
- Similarly, are the drug effect entities useful in identifying watchlist drugs?

This research is accomplished through an empirical study using a corpus from the Yahoo! public health forums, against which Chee applies various NLP techniques to define and distill a feature space for classification using Niave Bayes and Support Vector machines for detecting watchlist drugs. Drugs detected are evaluated against watchlist drugs found via the FDA Adverse Event Reporting System (AERS) to determine the utility of the approach and its applicability in Pharmacovigilance.

1.1 Background on Pharmacovigilance, AERS and Social Media

The dissertation begins with an extensive background discussion on adverse drug reactions and current surveillance techniques. Adverse drug reactions are defined by the FDA and World Health Organization (WHO) as "A response to a drug which is noxious and unintended and which occurs at doses normally used in man for prophylaxis, diagnosis, or therapy of disease or for modification of physiological function."**?**. Chee continues by introducing Pharmacovigilance as "the study of drugs once released to market"Chee, and the important regulatory agencies practicing it are mentioned - the World Health Organization (WHO) and United States Food and FDA. The FDA Adverse Event Reporting system (AERS) is discussed as comparison with it is central to the work. AERS was constructed to house mandatory drug safety reports from drug manufacturers, distributors and health care facilities, as well as voluntary reports submitted by consumers (patients), physicians and other healthcare providers. Reports are evaluated by the Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER) within the FDA for drug safety signals, which may then be elevated for further review by clinicians, epidemiologists and other expertise to determine the next steps, up to and including the removal of a drug from the market.

Chee identifies a major limitation in AERS and other *spontaneous reporting systems* in that they are known to have high underreporting rates (**?**), due to the likelihood of a patient reporting an event only if they feel their healthcare provider has not paid attention to the adverse drug reaction observed (**?**). This deficiency is presented as motivation for Chee's work exploring social media as a data source. Social media provides a venue for patients to share their health information in anonymous setting as patients are not always transparent nor truthful with their physicians. Online health forums create an environment where patients can find those having similar backgrounds, conditions and challenges, which in turn prompt rich social interactions where patient disclose their opinions and observations about their current drug regimin effectiveness and perceived adverse events. Chee feels these forums represent an untapped means to crowdsource data for the pharmacovigilance task.

1.2 Experimental Data

The data selected for the dissertation's experimentation is a Yahoo! corpus containing 12.5 million messages from various Yahoo Health group forums. As the data is a raw export containing a combination of message metadata, raw text and HTML, it must first be studied to better understand its composition and what NLP techniques should be applied to better prepare it for experimentation.

Chee conducts an initial study of the Yahoo! corpus by selecting at random 500 messages, then developing lexicons to understand their composition. The messages first prepared by

1.3 Patient engagement in Social Media

FDA watch list candidate. These observations also serve as a basis for selecting Yahoo! public health message forums as the principal data set used in the dissertation's experimental study. As of May 25, 2008 there are 162,754 groups to draw from (?). 12.5 million messages were made available for the analysis.

The Yahoo! corpus contains - meta-data about messages that is largely useless (name of group, inception date, is it moderated, language of group, group type). Chee asserts this is useless as it largely changes over time. - machine learning techniques are trained on grammatically correct text - Chee developed a variety of lexicon to interpret the messages and provide Statistics - tokens not in lexicon are manually inspected as spelling errors, differentiating between medical and general terms, names, foreign languages, web terms (HTML artifacts), and numbers - messages were sampled from the

¡discuss more about the data here¿

1.4 Sentiment, Named Entities and Classification

Of specific interest to the dissertation is applying sentiment analysis and named entity recognition to the messages in these forums. Sentinment analysis is presented as challenging because the domain dependent nature (??) can make it difficult to differentiate between positive and negative sentiment on words and phrases alone. Chee's approach is to calculate the probability of a specific word given a positive or negative class: $P(word|negativeorpositive)$. The hand crafted lexicons Linguisitc Inquiry Word Count (LIWC) and SentiWordNet are leveraged to generate sentiment scores on words. Support Vector Machines (SVMs) trained on words as features can also be used to separate positive phrases of text from negative.

Named entity recognition (NER) is necessary for identifying drug names and effects, such as headaches or vomiting. The challenge posed by doing so on forum data is the relaxed structure and oft-present grammatical errors make leveraging existing NLP tools trained on grammatically correct text difficult. Chee draws upon the work of Hearst (?) for automatically acquiring hyponyms from text. Hyponyms are words having more specific meaning than general or subordinate terms, thereby providing strong indication the discovered words are drugs or drug effects.

Classification techniques are employed by Chee to solve the problems of NER, sentiment analysis and assigning class labels to the message forum text. Specifically, Support Vector Machines (SVM) and Naive Bayes classifiares are used.

1.5 Support Vector Machines

SVMs map features into a high-dimensional space using a kernel function (?). A hyperplane is constructed that defines the decision boundary between two classes in this decision space, with those new observations being classified based on which side of the hyperplane they fall on. Chee quotes studies by Forman Joachims stating SVM's strengths in text classification, justifying its use in comparison to Naive Bayes. - LibSVM was used with a radial basis function (RFB) kernel - SVM solves the following optimization problem

$$min_{u,b,\xi}\frac{1}{2}w^Tw+C\sum_{i=1}^{l}\xi i$$

$$y_i(w^T\phi(x_i)+b)1-\xi_i$$

$$\xi_i 0$$

- RBF's are non-linear in nature which gave some accuracy advantages over linear - RBF's are trained on two parameters C and γ - Grid search method using cross-validation is employed to look for C and $gamma$ because it parallelizes well - C is the penalty parameter for the error term - RBF kernel is defined as $K(x_i,x_j)=exp(-\gamma||x_i-x_j||^2),\gamma>0$

1.6 Naive Bayes Classification

The dissertation uses Naive Bayes classification to address the NLP problems faced by Chee. Their use was somewhat counterintuitive because Naive Bayes Classifiers assume independence of features (words), whereas we know in real world settings that if a word like ”aspirin” were present, there is a greater probability of the the words ”headache” or ”pain” being present than ”lemonade”. However, in applied settings they still do reasonably well (**??**). NB has done well in SPAM detection (**?**) and make sense as a first step for their simplicity (no hyperparameters).

- given word grams w in messages about a drug D - $p(w_i|C)$ probability the $i-th$ word is from class C, C is watchlist or non-watchlist drugs. - $p(D|C)=\prod p(w_i|C)$ - probability of a given drug given the class - W = watchlist, so $P(D|W)=\prod p(w_i|W)$. - Bayes rule writes this as

$$p(W|D)=\frac{p(W)}{pD}\prod p(w_i|W)$$

$$p(\neg W|D)=\frac{p(\neg W)}{p(D)}\prod p(w_i|\neg W)$$

Chee combines these two probability modles with the maximum a posteriori (MAP) decision rule to pick the most likely hypothesis. ¡discuss maximum a posteriori method¿ - The method of MAP then estimates θ as the mode of the posterior distribution of this random variable

1.7 Feature Selection

- BNS (Bi-Normal Separation) cited by (**?**) outperforms other methods for rating ranking feature importance for Classification

- IG (Information Gain) Best practice suggested words occuring less than 3 times in a data set should be removed

1.8 Evaluation Metrics Used

- watchlist drugs are the positive examples - non-watchlist drugs are the negative examples - watchlist drugs that are false positive are the interesting ones from Classification - had to work with a 90/10 split where 90% of instances are one class (non-watchlist) and 10% are another (watchlist) it is difficult to outperform a naive classifier that just marks everything as non-watchlist - Receiver Operating Characterisitcs (ROC) curves are used with their Area Under the Curve (AUC) evaluated. - ROC curves are the true positive

1.9 Experimental Process

defining the speration

- Yahoo data pre-processing techniques using NLP - Selection of General Vocabularies and lexicons of interest - measurement of sentiment in drug outcome messages - selection of test training sets - classification using various lexicons - how KL Divergence is used as part of the experiment 3

3 pages

2. Discussion of Contributions

Chee asserts that his research develops a technique for discerning drug Safety events from public data sources, and that he is developing a "crowd sourced" means of Pharmacovigilance. Is this the case? Specific aspects to discuss are:

1) Exploration of classification techniques to discern FDA watch list drugs from free text 2) Exploration of the Yahoo! public Health message forums as potential data ource for adverse drug event mining 3) Approach generalizes to other social mediums?

3. Research Critique

A discussion of the methodology - were there any gaps? - good parts/bad parts?

4. Literature Review

What else is out there that is relevant in this space? Other studies that have used public data for medical purposes?

5. Application Areas

The most likely area is drug safety surveillance in uncontrolled settings.

- could we also use this in discovering underlying conditions? - drug combinations? - confounding fator disccovery?? Write up 1 to 2 pages here.

6. Concluding Remarks

Conclude the critique with a few endcap statements about what I learned from it, where it could go, how it could motivate future research, etc.

7. Paper Criteria (Grading)

The critique should include a summary of the research reported, a discussion of the major contributions claimed, and an assessment of the significance of those contributions and of the research itself. The critique should also include a brief literature review of the topic related to the thesis, discussion of relevant algorithms, and application areas for the research reported.

Where appropriate, the critique should include a comparison with other issues discussed in class. Students are encouraged to select a dissertation that is related to their course projects. The evaluation criteria for the critique are as follows: Overview of the research reported (20 Review of the related literature (15 Major contributions of the thesis (20 Understanding of techniques and algorithms (20 Application areas (15 Proper construction and readability of paper (10

References