Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

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.

1.2.1 Tokenization Study of Data

Chee conducts an initial study of the Yahoo! corpus by selecting at random 500 messages, stripping them of html tags, numerical and punctuation only tokens ($, %, :), :(, etc), then tokenizing them on white spaces with trailing punctuation. The tokens are then evaluated

in several rounds of classifiction using lexicons obtained or constructed by Chee to aid in understanding message composition. Lexicons for English and foreign language were drawn from the OpenOffice project (**?**). Drug names were taken from the Drugs@FDA website. A medical and disease terminology lexicon was constructed from terms on MedicineNet, Wikiepedia, and the MedDRA lexicon from FDA AERS. The names lexicon was constructed using names extracted from email addresses in message headers in the corpus, popular baby names from the United States (US) Social Security Administration, and popular common names from the US 1990 Census.

The classification process was iterative. If a token did not initially classify as English, web (a lexicon of web slang), medical, or drug name it was manually inspected and classified into *error types*: Foreign Language, Names (augmenting name lexicon), Spelling Errors, Compound Words, Slang, Abbreviations, Web, Unknown Words, Numbers and Garbage. Chee's analysis produced some interesting average metrics for the messages:

- Average of tokens per message: 172.21
- Average of drug name tokens: .29
- Average of error type tokens: 7.09
- Average of name tokens: 5.34
- Average of medical tokens: .81

Of the error tokens, over 54% of them were found to be foreign language tokens - primarily Indonesian and Spanish - motivating Chee to incorporate Foreign Language lexicons from OpenOffice to speed up classification. A primary concern for Chee was the presence of spelling errors, but the classification results show only a .8% error rate, which Chee uses to rationalize sticking with dictionary based approaches for word classification for their high precision. Finally, it is acknowledged that Named Entity Recognition (NER) is challenging in this context. FDA approved drugs represent a closed class of nominals, but foreign drugs, herbs and other chemicals are not available in a comprehensive list. Dictionary approaches to classifying drug outcomes are challenged by the use of slang terms.

1.2.2 A Vocabulary for Experimentation

Chee describes performance concerns training SVMs for classification using all the words in the message, given the $O(kn)$ training time for n training instances using k features (words). Additionally, multiple words together in order can convey a different meaning than separate single words, such as "vitamin a" compared to "vitamin" and "a".

These constraints motivate the use of word-grams - unigrams, bigrams and trigrams specifically - as a way to capture more accurate meaning. Chee references (**??**)'s work proposing the most informative words in a message would be the mid-frequently occuring ones, electing to take the top $k - n$ most frequently occuring terms in a message as the most important terms, where k is the top number of terms, minus n accounting for simple function words like *a*, *or* and *the*.

Specialized lexicons are developed as a means to ensure the classifiers that will be trained do not overfit to only those drugs in the drug lexicon, preventing the identification of previously unseen (unlabeled) watchlist drugs. The lexicons selected to use in the classification are:

- drugs - a drug list from drugs.com
- medical - medical terminology extracted from MedicineNet
- sentiment - a sentiment lexicon from the combination of SentiWordNet and Linguistic Inquiry and Word Count (LIWC)
- medra - the MedDRA terms for drug adverse events/outcomes from FDA AERS
- disease - a disease list from Wikipedia

These five lexicons allow for twentynine different datasets of features to be constructed from the Yahoo! corpus messages for the watchlist drug classification experiments. The feature vector used in classification is then the intersection of those terms found in *all* of the lexicons used in that particular test. For example, if the sentinment and drug lexicons are used together, the feature vector has only those terms that occur in both lexicons.

1.3 Language Identification for Messages

The previous study identified a significant number of foreign language messages in the corpus. While these text processing techniques are language agnostic, removing the foreign language messages will reduce the feature vector length for training, as well as acknowledge the audience for this study is English speaking.

Messages containing non-romanized text are removed first using Unicode language detection. Since non-English languages in romanized text are harder to discern, Chee compares and contrasts character n-gram approaches by (**??**) with dictionary approaches from (**?**) for foreign language classification. Dictionaries are opted for given the simple, binary nature of the problem: Is a message English or not? Dictionaries for foreign language words are taken from the OpenOffice project and used in conjunction with the medical, drug, disease and name lexicons mentioned earlier are used to evaluate a linear inequality for each message to determine if it will be kept or not.

First the messages are stripped of tokens containing web addresses, punctuation only (emoticons), as well as short words, email addresses or tokens that are already on an ignore list. Remaining tokens in each message are counted up using the following algorithm:

- if (word in ignore) OR (word length ¡ 2) OR (word contains "@") ++ignore
- else if (word in English) ++english
- else if (word in Drug) ++drug
- else if (word in medical) ++medical
- else if (word in name) ++name

- else if (word in foreign) ++foreign
- else unknown++

Once the counts are obtained, the following linear inequality is evaluated:

$$4 * foreign + unknown + ignore > english + drugs + medical$$

The weighting of the foreign words is selected to ensure messages contain less than 25% foreign words to be considered english. Foreign messages are removed and English messages retained. This resulted in a reduction from 12,520,438 messages to 10,178,710 messages, and a reduction in the number of unique terms per message from 2.5 to 2.

1.4 Experimental Design

The goal

- feature vector 1 is entirely words or phrases, including N-grams where word ordering matters, lexicons used - second vector is words from vector 1, then speciality features: counts of instances of a speciality lexicon. Number of diseases mentions, drug mentions, positive sentiment words, negative sentiment Words -utilized Kullback-Leibler divergence (KL divergence) to quantify the difference in word frequency distributions between watchlist and non-watchlist drug containing messages - created a probability distribution from a frequency distribution over words utilizing smoothing (smoothing handles the case of a word in P not in Q creating infinite divergence) - compuared KL divergence between watchlist and non-watchlist drug containing messages to the Google Web 1T 5-gram corpus and the Reuters Corpus. - google corpus is english word n-grams from 1 to 5 words, from 1 trillion words on public web pages (used only single words in study) - Reuters is english language articles from 8/20/97 to 8/19/97 (810,000 articles) - Hypothesis is that the KL Divergence between the Yahoo! corpus and Goole is lower because of the colloquial nature (brand names, slang, proper nouns, etc) and Reuters is more formal (good grammar, etc)

Dictionary Named Entity Recognition (NER) - dictionary approach is used due to lack of training data NLP tools if we take a statistical classification approach - drug lexicon developed earlier is used with names from FDA drug taxonomy on Drugs.com. The taxonomy allows group of name brand and generic drugs by function. Allows grouping drugs to eliminate data proverty. - AERS uses MedDRA. Medical Dictionary for Regulatory Activities (MedDRA). Used to report AEs from clinical trials. - Lucene index created over the 10million messags. AERS and drug lexicons used to generate phrase searches, run against the Lucene index. - limitation: using a lowercase filter with stemming. Creates a lowercase representation of a term within the index: Commit and COMMit are "commit". "committing" is mapped to "commit". Implications for precision and recall. - lower casing results in greater recall (of documents found) w/ less precision (of relevant documents from those found). Drugs containing common words like "Control" present a problem. We also lack contextual clues or part of speech (POS) tags to further identify if the drugs are present. - FUTURE area: IR (Information Retrieval) system could be augmented to use context such as fuzzy searches requiring "drug" or "taking". POS tagging could help as well - differentiate between verb commit and noun Commit. For purposes of this dissertation,

drug names with common words were replaced by the generic (chemical) name. - For each lexicon phrase, searches were constructed and run against the index.

- Concerned about distance between unique drug mentions within messages (more than 1 drug). If drugs are close together it is harder to segment the message to determine sentiment and effect regarding a specific drug.

- pairs of unique drug mentions within the same message were extracted and the distance (number of characters) from the first mention are looked at. - drawback: messages containing more than 2 unique drugs. Eliminating this by looking at messages with 2 mentions (??) - frequent co-occurances between different drugs highlights how the drugs are talked about - variance in distance illustrates that drugs are talked about in the same way - a big difference between median and mean indicates variability in the data - Mutual information used to score pairs of unique drugs with the same message to look for highly co-ocurring drugs (this was used as a scoring function instead of cosine similarity)

- Sentiment is reflected in 2 features on the word feature vector: number of positive sentiment containing terms number of negative - Does the sentinment actually reflect the authors opinion of the drug? - Look @ aggregate sentiment over time. Changes in sentiment reflect news about the drug. - Drugs may have serious side effects, yet people may still have a positive attitude towards it if it helps them in some way. - Method uses portions of the lexicon in the LIWC when calculating sentiment scores for messages. References previous work by (**????**) to demonstrate variations in language usage between depressed and depression vulnerable students. - Used the words in LIWC corresponding to: positive emotion, negative emotion, anxiety, anger and sadness. Augmented LIWC lexicon to include a wide range of emoticons: :), :(, :P, $^{\cdot LOL, ROFL. - Messages from Yahoo groups were reparsed to extract just the textual information. and remove replies. - Negative valer}$

Limits of Classification - few watchlist drugs - only 435 drugs w/ more than 500 unique messages mentioning them, only 63 of which are watchlist drugs. - reducing to 250 unique messages, we have 575 drugs of which only 77 are watchlist - 90% of instances are non-watchlist - Cross-fold validation is utilized to alleviate some of the data sparsity problem - This dissertation uses random undersampling (RUS) to balance out the distribution of positive and negative samples. The majority class (negative) is randomly discarded. RUS resulted in the best performance in empirical tests of 2340 data sets. Area to comment on: Other techniques?

- Decision Making. - Scaling of individual dimensions has to be considered.

1.5 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.

1.6 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

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

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

$$\xi_i 0$$

- 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 $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2), \gamma > 0$

1.7 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

1.8 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

1.9 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

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