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

Table 1:

- 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 Experimentation and Results

The goal of the dissertation is to develop a classification system for drugs based on how people are talking about them in online message forums. Processed forum messages are first organized by drug, divided into test and training sets, converted into feature vectors, and then run through Support Vector Machine and Naive Bayes classification algorithms. Chee conducts a multitude of these experiments using two versions of the feature vector structure, multiple combinations of the lexicons described earlier, usage of top $n - k$ terms and BNS for term selection to feed feature identification using the lexicons, as well as cost-weighted and unweighted variants of Naive Bayes and SVMs.

1.4.1 Class Separability Confirmed via Kullback-Leibler Divergence

Chee begins with an initial experiment to determine how separable conversations regarding watchlist and non-watchlist drugs will be in order to validate the classification approach. Kullback-Leibler divergence (KL divergence) is used to measure the difference in word frequency distribution between messages in each category, with a smoothing technique applied to make sure each term is represented - albiet with very low frequency - in each distribution to prevent infinite divergence. A series of comparisons are done between between the watchlist and non-watchlist frequencies, as well as term frequencies in the Google Web 1T 5-gram corpus and Reuters Corpus to provide perspective.

- $D_{jk}(Watchlist||Non) = .1684$
- $D_{jk}(Non||Watchlist) = .1778$
- $D_{jk}(Watchlist||Google) = 1.4178$
- $D_{jk}(Watchlist||Reuters) = 1.3279$
- $D_{jk}(Non||Google) = 1.1804$
- $D_{jk}(Non||Reuters) = 1.0815$

Table 2:

Clearly Watchlist Nonwatchlist diverge, and show a different degree of divergence from the Google and Reuters corpi, indicating some separability.

Interesting term frequency differences:

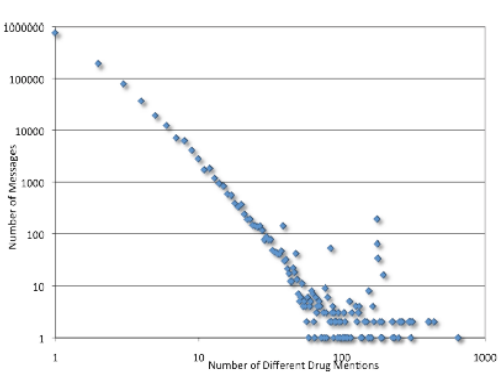
- i $D_{kl}(Watchlist||Non) = .007875$
- my $D_{kl}(Watchlist||Non) = .003018$
- me $D_{kl}(Watchlist||Non) = .0002256$
- you $D_{kl}(Watchlist||Non) = .001712$
- i'm $D_{kl}(Watchlist||Non) = 9.64E-04$

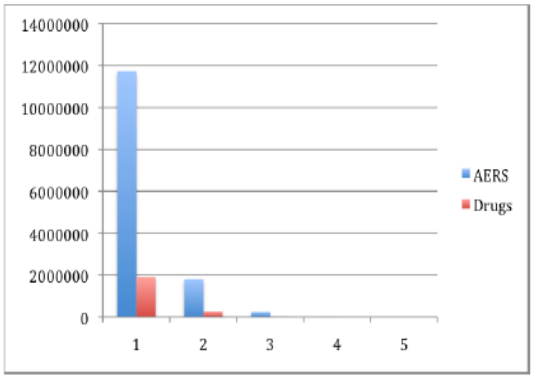
The above terms are overexpressed in the watchlist messages compared to non-watchlist, and are consider indicative of emotional writing - just the kind of writing that might occur when discussing a drug with an adverse effect. Chee concludes these results support separability of the two classes of message based on their word features.

1.4.2 Named Entity Recognition in Messages

Another challenge presented by Chee is how to identify the drug(s) and adverse event(s) themselves within a message - a problem of Named Entity Recognition (NER). A dictionary based approach using a drug lexicon compiled from FDA and Drugs.com, and adverse event lexicon using the Medical Dictionary for Regulatory Activities (MedDRA) is used to query a Lucene index built atop the processed Yahoo! corpus messages. The index construction applied a lowercase filter with stemming, presenting a problem with common words that are also part of drug names. For example, the drug name *Commit* is indistinguishable from the verb *commit*. This was addressed by replacing drug names with common words by their generic (chemical) name in the lexicon to favor precision of query results over recall.

A series of phrase searches were issued against the index to determine prevalence of MedDRA terms (adverse events) and drug names. Two important conclusions are drawn thanks to the resultant plots:





The first plot demonstrates a zipfian distribution for drug name mentions in the messags, with more than 96% of messages mentioning at most 5 drugs, adding confidence to the

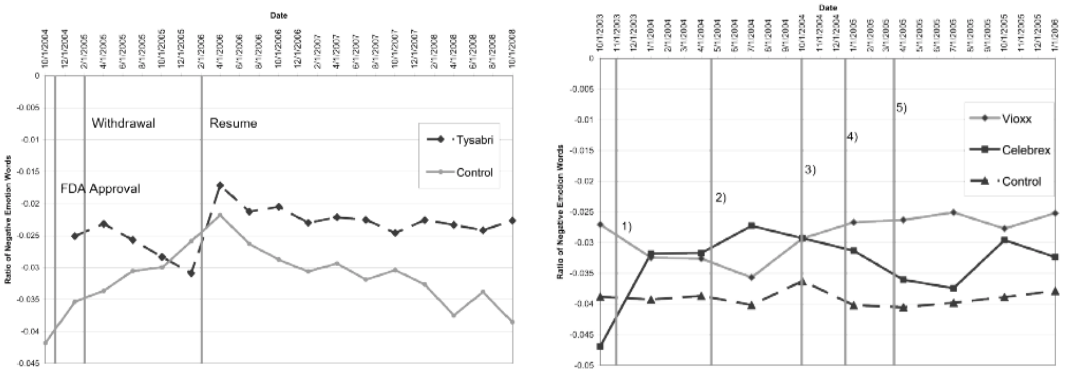
hypothesis that messages can be associated with at most a handful of drugs. Chee elects to eliminate messages having more than 5 unique drugs as these frequently constitute lists of drugs posted or SPAM messages. The second plot shows that adverse events are mentioned far more often than drug names, suggesting that a preceeding post containing a drug name eliminates referencing it in subsequent posts - much like replies on a forum. Investigating this is an area for future research. Note that the numbers on the X-axis for the 2nd plot refer to the number of terms (1, 2, ...) that comprise the drug name or adverse event - a fact not really relevant for this analysis.

1.4.3 Handling Multiple Drug Names

If multiple drugs are mentioned in a single message, it becomes difficult to discern which drug the message should apply to. Another study is conducted to evaluate the typical distance (number of characters) between the top 25 co-occuring drugs in messages to determine if the messages could be segmented into relevant portions for each drug. The analysis shows a Zipfian distribution in the character separation, indicating that most drugs are talked about together within a single sentence, or in adjacent sentences, leading Chee to conclude separation is not possible and that any adverse events mentioned in a message should just be attributed to all drugs mentioned within the message.

1.4.4 Sentiment Feature

A final experiment preceeding the main classification work is done to validate how message sentiment will be determined. Chee hypothesizes that the positive or negative valence in a message represents drug satisfaction, making it an interesting feature to incorporate into the classification experiments. A lexicon using the positive emotion, negative emotion, anxiety, anger and sadness terms from LIWC is constructed, augmented to include several emoticons (:), :(, ..) and acronyms (LOl, ROFL, ..). Two case studies are then executed using messages sampled from specific groups in the corpus, and the change in drug sentiment is analyzed over the drug's pre-recall, recall, and post-recall timesframes. This sentiment is compared with a control sentiment derived from those messages in the samples not containing the drug to look for a statistically significant difference.



The left-most plot shows the sentiment change for Tysabri pre-recall, recall (withdrawal) and post-recall (resume) over the control. It shows a reasonably intuitive change in negative valence to Tysabri having been introduced (more positive), withdrawn (negative), reintro-

duced (hopeful therefore positive), then stabalizing. The right plot shows sentiment change for a pair of commonly used pain relievers - Vioxx and Celebrex - over the course of several public anouncements and a withdrawal of Viox (sections 1 through 3), then Celebrex (4 and 5). ANOVA is applied to both case studies to determine statistical significance for each drug in each segment against the control. Both were found to be statistically significant with p .001.

1.4.5 Feature Selection, Training and Test Data Size

The introduction to the main body of classification experiments is preceeded by a brief discussion on the features vectors used, as well as how training, test and validation sets are constructed to support the use of 10-fold Cross Validation for classifier evaluation.

Two types of feature vector are decided upon and then leveraged in the experiments. The first feature vector type is generated over general vocabulary terms in the messages, selected based on frequency cutoffs. This vector is then augmented with counts from the various specialized lexicons mentioned earlier: medical, diseases, drugs, sentiment and reactions (MedRA). The second feature vector uses only the specialized lexicon.

The richness of the training data is of foremost concern to Chee. There are only 435 drugs having 500 or more unique messages, and only 575 drugs having more than 250 messages, with 63 and 77 watchlist drugs mentioned in each respectively. Therefore approximately 90% of message instances reference non-watchlist drugs, creating a data scarcity problem when attempting to classify watchlist drugs. Chee decides upon a minimum cutoff of 250 messages per drug for that drug's messages to be included in training and testing. An experiment is constructed to evaluate techniques to address the scarcity, including scaling features, selecting different ratios of negative to positive training examples such as 1 to 1 and 2 to 1, and experimenting with different split ratios in cross-validation - 90/10 and 80/20. These experiments are dubbed inconclusive and not elaborated on further in the dissertation.

1.4.6 Classification Experiments and Results

- test and training sets are sampled with the same distribution as the original data - 90/10 split, 90% to train and 10% to sample. Each split is sampled to preserve the original distribution of positive and negative instances.

- cost weighting and no-cost weighting experiments are run. cost weighting adjsuts the costs of the positive and negative examples so they are approximately equal. A greater penalty is imposed for incorrectly classifying a positive example than a negative one.

- Decision Making. - Scaling of individual dimensions has to be considered. Avoid attributes having greater numeric ranges dominating those with lower. Term-frequency inverse document frequency (tf-idf) can be used to weight a word according to its importance upon the number of appearances.

- 90/10 split for training presents a porblem because if only 10% of the instances were used for testing, this leaves 58 testing instances or 16 instances with a balanced test set.

- Chose to use drugs with more than 250 messages, leading to an overall example set of 575 drugs, of which 77 are watchlist drugs. Separating examples into two different data sets - one with 500+ and one with 250-500 leads to questions of generalizability. Will a

classifier trained on more data (500 messages/drug) work on instances with sparser data (250 messages/drug) or vice versa? Would the results be fair? Both types are mixed and a subset of the total is used for testing while the rest is used for training - 50% 500+ - 50% 250-500

An initial experiment was run to answer some data set construction questions. - should we use a 1:1 ratio for testing, or 2:1 - chose an 80/20 split since if a 90/10 was used with 1:1 we would have only 15 positive instances - word features used were comprised of the combined special lexicons - Used SVMs with RBF and grid searh with 10-fold cross-validation for building models - test data was run, ROC curves generated and AUC Numbers - These tests were inconclusive

For the classification experiments training sets are sampled with the same distribution as the original data. Data is divided into a 90/10 split with 90% samples being used to train and 10% used for testing, and are sampled such that the splits represent the original distribution of positive and negative instances. - several experiments are run. First a general classification experiment with now cost weighting. Next a cost weighting experiment where the costs of positive and negative examples are adjusted to make the distributions equal. In the case of cost weighting experiment, a greater pentalty is imposed for incorrectly classifying a positive example then a negative one.

- types of experiments performed:
beginitemize

Un-normalized Naive Bayes (UNB)

Un-normalized Naive Bayes w/ cost weighting (UNBC)

Normalized Naive Bayes (NNB)

Normalized Naive Bayes with cost weighting (NNBC)

Unnormalized SVM (SVM)

Unnormalized SVM with cost weighting (SVMC)

Normalized SVM (SVM)

Normalized SVM with cost weighting (NSVMC)

Grid search was performed for each fold on SVM experiments to tune the hyperparameters.

Speciality Lexicon selections - initial experiment of the 5 combinations (medical, disease, drug, sentiment, reaction) was used to identify the best combination of lexicon that provides top performance. 240 experiments with various parameters were run. Accuracy, F1 score and AUC were chosen as evaluation metrics.

BNS lexicon experiments. - Use test subset of data to chose the most salient word gram features using Bi-Normal Separation which preferences wordgras that are differentially expressed between watchlist non-watchlist messages. The top 15,000 , 10,000 and 5,000 word grams were chosen from the test subset. Utilizing word grams from the entire message would inflate the classification scores since often times classifiers do not have complete knowledge about the entire lexicon apriori.

Watchlist predictions - previous sections results are used to choose the highest performing classifiers for Accuracy, F1 and AUC scores. Using the feature set and classification

algorithms, numerous cross validation runs across all of the data are performed multiple times. The output from these classification runs provides insight into future watchlist drug predictions.

- multiple top performing classifiers from each category are combined in an ensemble like approach to produce a meta-classifier where the false positives from each category are combined using a linear combination resulting in a score

- hypothesis is that drugs consistently marked as false positives are candidates for watchlist drugs in the future.

- Given a 90/10 split where 10% of drugs are used to evaluate a classifer, 10 runs should ensure each drug is tested at least once and 50 runs statistically speaking should allow each drug to be classified 5 times against 5 different classifiers. For these experiments, each set of features was used to build a hundred classifiers, test and training sets.

- Outputs from classifers are utilized because different types of classifers were found to perform the best and their output cannot be directly compared.

- a weighted ratio is creatd to score the false positives

$$\frac{NumberofFalsePositives}{NumberofOccurences(tests)} * NumberofFalsePositives * NumberofClassifierTypes$$

Produces a weighted averaged over the number of false positives. The number of different classifiers is also important, which classified it as a watchlist drug giving credence to other classifications.

Two runs were made with drugs withdrawn from the market. First widthdrawn drugs were labeled as non-watchlist to determine if the classifiers would accurately identify the withdrawn drugs. Second run removed the watchlist drugs and classifies them after the classifier has been built for each fold of the cross-validation run. This second method should more accurately identify the withdrawn drugs with greater confidence because their data is not mixed with the other non-watchlist drugs possibly, reducing the accuracy of the classifiers.

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 spe-

cific 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