Evaluating the Adverse Drug Reactions Learning Framework Using Novel Text Mining

K.M. Deepa Lakshmi, Dr. A. Prakash

1 Research Scholar, 2 Professor,
1&2 PG & Research Department of Computer Science,
1& 2 Hindusthan College of Arts & Science,
1& 2 TamilNadu, India.

Abstract

Drugs are used to cure disease, but most of the drugs in the markets produce adverse drug reaction commonly known as side effects. This Adverse Drug commonly led to injury or death during medical treatments for patients. Only very few Adverse Drug Reactions are identified before a drug is marketed. Most drug adverse reactions are identified only after its intervention in the market. The post-marketing drug observation methods are based on the reports provided by the history of patient records and recommendations of doctors, this result in loss of identification of the earliest indications necessary to prevent the occurrence of such injuries or deaths. There are few online healthcare forums which provide some prior information to medications. The overall objective of this research is to extract reports of adverse drug side-effects from the information’s available in online healthcare forums and use them as early indicators to assist in post-marketing drug observation. In this research we follow a novel method of extracting adverse side-effects of drugs from health care forum messages as a sequence labeling problem and present a Novel Hidden Markov Model (NHMM) based Text Mining system that can be used to classify a message as containing drug side-effect information. This is further extracted with the adverse side-effect mentions from it. This common forum is used in the training and validation of the NHMM based Text Mining system. The outputs of the system are consolidated and finally interpretation from the results is useful by gathering the pre medication results.

Keywords: Adverse Drug Reaction, Side effects, Text Mining, Novel Hidden Markov Model

INTRODUCTION

Generally medicines or Pharmaceutical drugs are product of chemical combination of substances in a certain mixture prescribed for the medical treatment or cure of diseases and other health disorders. A side-effect is an unpremeditated response that is experienced by a patient due to the intake of a drug. Side-effects can be of two forms either positive or negative; however, it is the negative side-effects or Adverse Drug Reactions (ADRs) that are more important, as they can severely affect the health of patients, sometimes fatally. By leaving out the laymen, even doctors don't get adequate information about the drugs they prescribe to their patients. Companies advertise prescription drugs in medical journals to boost sales, but a large majority of them don't give doctors vital information such as adverse effects the medicines can have on patients, according to a study published in a recent edition of Indian Journal of Medical Ethics.

In India it is estimated that over million serious ADRs occur among hospitalized patients, which results in over huge deaths each year making ADRs a significant public health problem. Drugs are approved for use by the general public only if their therapeutic effect outweighs their adverse side effects. Drug manufacturers are mandated to publish the side-effects that have been identified as a part of the clinical trials. These are usually published as a part of the Drug Package Inserts or Drug Package Labels for each drug. However, the clinical trials are often not extensively enough to uncover all possible side-effects due to the small number and diversity of the participants involved. In order to address this issue, health organizations around the world employ post-marketing surveillance programs as a part of their Pharmacovigilance: the science relating to the detection, assessment, understanding and prevention of adverse effects of pharmaceutical drugs. All the reported adverse events are recorded as a part of the FDA Adverse Event Reporting System (FAERS) and are constantly monitored for statistically significant adverse drug event reports. Once such reports are confirmed against a drug, the FDA may take necessary action against the drug manufacturer, sometimes by completely recalling the drug from the market. However, with the spontaneous reports being purely voluntary, not all adverse events get reported.

It could take several years before a significant number is reported to initiate an inquiry, analysis and follow up action, during which, the drug could continue to affect a larger percentage of the general population. Thus, there is a need for systems that can help in the early detection of such Adverse drug events.

Methods for automatic extraction of adverse drug events can be categorized based on the nature of the data sources: structured and unstructured. The spontaneous adverse event reports collected by the health authorities are the major sources of structured data, which, though varying in format, are suitable for data mining. Reviews on data mining algorithms that have been used to extract adverse side-effects of drugs from such structured data sources are discussed in . Information on adverse reactions of drugs is also widely available as a part of unstructured data sources such as: literary sources like to publish biomedical literature, including
books, journals and papers, along with clinical sources like patient medical history and online healthcare forums.

**Objective** –

The objective of this research paper is to extract reports of adverse drug side-effects from messages in online healthcare forums and use them as early indicators to assist in post-marketing drug surveillance. The objective is achieved through Novel Text mining mechanisms such as Hidden Markov Model.

**Methodology** –

The task of extracting adverse side-effects of drugs from health care forum messages as a sequence labeling problem and present a Novel Hidden Markov Model (NHMM) based Text Mining system that can be used to classify a message as containing drug side-effect information and then extract the adverse side-effect mentions from it.

**HMM Text Mining Algorithm** -

A hidden Markov model, or HMM, is a particular kind of probabilistic model based on a sequence of events—in terms of token identification, this represents sequential words in text. Although different approaches have been studied and implemented, the best-known token identification system that incorporates a machine learning component is based on hidden Markov models. Identifinder is a well-known system. It uses a variant of a hidden Markov model to identify tokens like names, dates and numerical quantities. Each state of the HMM corresponds to a token class. There is a conditional state for “not a token class”. Each individual word is assumed to be either part of some predetermined class or not part of any class. According to the definition of the task, one of the class labels or the label that represent “none of the classes” is assigned to every word. Identifinder uses word features, which are language-dependent, such as capitalization, numeric symbols and special characters, because they give good evidence for identifying tokens.

Identifinder (Bikel et al., 1999) is a well-known system. It uses a variant of a hidden Markov model to identify tokens like names, dates and numerical quantities. Each state of the HMM corresponds to a token class. There is a conditional state for “not a token class”. Each individual word is assumed to be either part of some predetermined class or not part of any class. According to the definition of the task, one of the class labels or the label that represent “none of the classes” is assigned to every word. Identifinder uses word features, which are language-dependent, such as capitalization, numeric symbols and special characters, because they give good evidence for identifying tokens.

**TEXT EXTRACTING PROCESS**

Text Mining systems are primarily used in the discovery and extraction of knowledge from unstructured text data. Figure 1 presents the architecture of our Text Mining system used for extracting Drug-Side Effects relationships from online healthcare forums. It primarily consists of the following 3 Process:

- Information Retrieval Process to create a collection of relevant documents
- Text Processing Process to preprocess text in the collected documents to facilitate extraction
- Information Extraction Process to extract information of interest from preprocessed texts

**Information retrieval Process**

The Information retrieval Process consists of a system that is responsible for extracting relevant documents or data sources from which we are interested to extract useful information. Some of the common approaches for data collection include: collecting results from search engines, the dataset from the medications.com was primarily used in the training and validation of the HMM classifier, while the one from steadyhealth.com was used in the analysis of the mined side-effects. It is to be noted that this study did not involve any experimental research on humans or animals; hence an approval from an ethics committee was not applicable in this regard. The data collected from the online healthcare forums are publicly available data and no personally identifiable information of the forum users were collected or used for this study.
Text processing Process

The Text Processing Process is used to extract textual units from document collections and process them into a format suitable for use by the Information Extraction Process. Typically, this Process is comprised of several Natural Language Processing (NLP) tools linked together as a pipeline for processing text data. Figure 1 presents the text processing steps in our system. In order to have a robust system that is not affected by the semantics of the language, we do not include techniques like part-of speech tagging, stemming or word sense disambiguation.

First, the crawled web document collection is parsed to extract unique thread names and associated messages. Each of these messages are then processed to remove HTML tags, converted to lower case and run through filters to remove unwanted punctuation and raw numerical data. The resulting text is then tokenized, filtered of common stop words and substituted with respective lexicon identifiers for ease of processing in the information extraction stage.

Figure 2: Process flow of Information Extraction Process

Figure 2. Text preprocessing and information extraction
Information Extraction Process

The Information Extraction Process is used to identify entities of interest in the preprocessed data and extract possible relationships between them. It consists of the Named Entity Recognition and Relationship Extraction sub-process. Named Entity Recognition helps to identify entities of interest in a given text. In our scenario, the entities of interest would be names of drugs, terms denoting side-effects and keywords or phrases that indicate a relationship between the drug and a side-effect.

When using the lexicon-based methods for performing NER, the choice of vocabulary that is used to create the dictionary entries has a significant impact on the performance of the NER Process. So it becomes necessary that the vocabulary of the dictionary reflect the vocabulary of the target corpus to be mined. The Relationship Extraction Process is used to identify the presence of relationships between the named entities in a given text. In general, several techniques including rule-based, statistical co-occurrence and natural language processing methods have been employed for this purpose. We make use of Hidden Markov Model (HMM), a supervised machine learning approach, to predict the presence of a relationship between a drug and an adverse side-effect.

If a message contains only a drug name and side-effect mention, it is not sufficient to denote a positive ADR. There needs to be some form of causal relationship that clearly associates the drug with the side-effect. It is in this regard that the keywords identified by the HMM are used to capture the causal relationship. As a part of the training, the HMM is trained on positive samples where it learns the association between the drugs and side-effects through the presence of keywords and uses this information for relationship prediction on the test data set.

NOVEL HIDDEN MARKOV MODEL

A Hidden Markov model is a statistical model in which the system being modeled is assumed to be a Markov process with hidden states. The outputs of the hidden states are observable and are represented as probabilistic functions of the state. In general, a HMM is defined using the following parameters:

- $N$: Number of states in the HMM
- $M$: Number of observation symbols in the HMM
- $A = [a_{ij}]: N$ by $N$ state transition probability matrix
- $B = b_j(m): N$ by $M$ observation probability matrix
- $\Pi = [\pi_i]: N$ by $1$ initial state probability vector

HMMs have primarily been used to model sequence data like speech utterances in speech recognition and Part-of-Speech tagging. They have also been successfully used for Information extraction and Named Entity Recognition. The success of HMMs in these tasks has motivated us to explore the possibility of using them to perform Relationship Extraction. The Hidden Markov Model library was used for implementing the Hidden Markov Model. This is an example taken from online health forum http://www.medications.com/bactrim

To extract relevant pieces of information from web pages, focused web crawlers are used at the target website. A Text Parser and forum message Extraction is done, finally Text processing is done.

Sulfamethoxazole and trimethoprim are both antibiotics that treat different types of infection caused by bacteria. The combination of sulfamethoxazole and trimethoprim is used to treat ear infections, urinary tract infections, bronchitis, traveler’s diarrhea, and Pneumocystis carinii pneumonia. Sulfamethoxazole and trimethoprim may also be used for other purposes not listed in this medication guide. Take this medication exactly as prescribed by your doctor. Do we can calculate the hidden states chain, based on the observation chain and using classification algorithm like viterbi algorithm or counter algorithms of hmm, one can find the most likely result.

Here the side effect identified is depression, which has been notified in S.No 9 & 11, in Table 5—Comparison of mined drug adverse reactions with those reported in drug package inserts.

EXPERIMENTAL RESULTS AND DISCUSSION

In order to compare the performance of the classifiers we do a 5-fold cross-validation on the 1000 sample manually annotated training dataset. Table 2 presents the results of a single run of the 5-fold cross-validation for both the Baseline and NHMM classifiers. The experiments are simulated using MATLAB under core 2 Duo 2.4 Ghz Machine. The results are compared with the general baseline classifier and Novel Hidden Markov model classifier. The below mentioned results indicate that the novel model produces a better result when compared to existing classifier in terms of its Quality of Service parameters such as True Positive, True Negative, False Positive, false negative, Precision, recall, F-Score and accuracy.
Table 1. General Classifier Results on Quality of Service parameters such as True Positive, True Negative, False Positive, false negative, Precision, recall, F-Score and accuracy.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Train set</th>
<th>Test set</th>
<th>True Positive</th>
<th>False positive</th>
<th>True negative</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1291</td>
<td>211</td>
<td>29.0</td>
<td>10.0</td>
<td>135.0</td>
<td>26.0</td>
</tr>
<tr>
<td>2</td>
<td>1291</td>
<td>211</td>
<td>25.0</td>
<td>16.0</td>
<td>137.0</td>
<td>19.0</td>
</tr>
<tr>
<td>3</td>
<td>1291</td>
<td>211</td>
<td>36.0</td>
<td>13.0</td>
<td>125.0</td>
<td>22.0</td>
</tr>
<tr>
<td>4</td>
<td>1291</td>
<td>211</td>
<td>28.0</td>
<td>14.0</td>
<td>130.0</td>
<td>25.0</td>
</tr>
<tr>
<td>5</td>
<td>1291</td>
<td>211</td>
<td>29.0</td>
<td>8.0</td>
<td>130.0</td>
<td>29.0</td>
</tr>
</tbody>
</table>

Figure 3: General Classifier Results on Quality of Service parameters such as True Positive, False positive.

Figure 4: General Classifier Results on Quality of Service parameters such as True Negative, False Negative.
Table 2. General Classifier Results on Quality of Service parameters such as Precision, recall, F-Score and accuracy.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Train set</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1291</td>
<td>0.7671</td>
<td>0.5090</td>
<td>0.6140</td>
<td>0.8270</td>
</tr>
<tr>
<td>2</td>
<td>1291</td>
<td>0.6250</td>
<td>0.5680</td>
<td>0.5950</td>
<td>0.8270</td>
</tr>
<tr>
<td>3</td>
<td>1291</td>
<td>0.7140</td>
<td>0.5140</td>
<td>0.7600</td>
<td>0.8160</td>
</tr>
<tr>
<td>4</td>
<td>1291</td>
<td>0.6590</td>
<td>0.5190</td>
<td>0.5810</td>
<td>0.8010</td>
</tr>
<tr>
<td>5</td>
<td>1291</td>
<td>0.8570</td>
<td>0.4910</td>
<td>0.5960</td>
<td>0.8060</td>
</tr>
</tbody>
</table>

In general our interpretation of the results is, the NHMM based classifier performed better with an average F-Score of 0.8 in comparison to the Baseline classifier which yielded an average F-Score of 0.6. It is evident that the Baseline classifier performs poorly in comparison to the NHMM classifier as both its the False Positive and False Negative values are higher. The higher False Negatives can be attributed to the fact that the baseline classifier is not able to
predict ADR relationship for drug/side-effect combinations that it has not seen before. The NHMM-based classifier, in contrast, is able to predict such relationships, even in cases where positive ADRs between a specific drug and its side-effect were not available as a part of the training set. It is in this regard that the NHMM classifier is capable of extracting some novel drug/side-effect information as well.

Table 3. NHMM Classifier Results on Quality of Service parameters such as True Positive, True Negative, False Positive, false negative

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Train set</th>
<th>True Positive</th>
<th>False positive</th>
<th>True negative</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1291</td>
<td>42.0</td>
<td>10.0</td>
<td>151.0</td>
<td>8.0</td>
</tr>
<tr>
<td>2</td>
<td>1291</td>
<td>39.0</td>
<td>12.0</td>
<td>140.0</td>
<td>19.0</td>
</tr>
<tr>
<td>3</td>
<td>1291</td>
<td>47.0</td>
<td>9.0</td>
<td>131.0</td>
<td>14.0</td>
</tr>
<tr>
<td>4</td>
<td>1291</td>
<td>29.0</td>
<td>17.0</td>
<td>146.0</td>
<td>7.0</td>
</tr>
<tr>
<td>5</td>
<td>1291</td>
<td>35.0</td>
<td>11.0</td>
<td>138.0</td>
<td>18.0</td>
</tr>
</tbody>
</table>

Figure 7: NHMM Classifier Results on Quality of Service parameters such as True positive, False positive

Figure 8: NHMM Classifier Results on Quality of Service parameters such as True Negative, False Negative
Table 4: NHMM Classifier Results on Quality of Service parameters such as Precision, recall, F-Score and accuracy.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Train set</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1291</td>
<td>0.7160</td>
<td>0.8510</td>
<td>0.8330</td>
<td>0.9180</td>
</tr>
<tr>
<td>2</td>
<td>1291</td>
<td>0.7710</td>
<td>0.6850</td>
<td>0.7250</td>
<td>0.8570</td>
</tr>
<tr>
<td>3</td>
<td>1291</td>
<td>0.8460</td>
<td>0.7720</td>
<td>0.8070</td>
<td>0.8930</td>
</tr>
<tr>
<td>4</td>
<td>1291</td>
<td>0.6440</td>
<td>0.7840</td>
<td>0.7070</td>
<td>0.8780</td>
</tr>
<tr>
<td>5</td>
<td>1291</td>
<td>0.7910</td>
<td>0.6800</td>
<td>0.7310</td>
<td>0.8710</td>
</tr>
</tbody>
</table>

Figure 9: NHMM Classifier Results on Quality of Service parameters such as Precision, Recall

Figure 10: NHMM Classifier Results on Quality of Service parameters such as F Score, Accuracy

Co-occurrence of a drug name and a side effect does not necessarily imply presence of a positive ADR. It is for this reason the False Positives for the Baseline classifier are higher. There needs to be a clear indication of a causal relationship that shows a drug is responsible for a side-effect. It is in this regard that the additional keyword information...
used by the NHMM classifier is capable of identifying the causal relationship between the drug and the side-effect. The False Positives in case of the NHMM classifier were identified to be caused primarily due to the lack of distinction between the symptoms that a drug is treating and the side-effects it causes. It can be addressed this by maintaining a list of symptoms for which a drug is prescribed and eliminate them from the list of side-effects identified to improve the accuracy of the classifier.

Table 5: Comparison of mined drug adverse reactions with those reported in drug package inserts

<table>
<thead>
<tr>
<th>S.No</th>
<th>Side Effects</th>
<th>% identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High blood pressure Headache, dizziness, cough, Cough</td>
<td>(12.57%)</td>
</tr>
<tr>
<td>2</td>
<td>dizziness</td>
<td>(2.77%)</td>
</tr>
<tr>
<td>3</td>
<td>Hearing loss</td>
<td>(0.53%)</td>
</tr>
<tr>
<td>4</td>
<td>fatigue, rash, diarrhea, nausea, headache</td>
<td>(1.81%)</td>
</tr>
<tr>
<td>5</td>
<td>shingles</td>
<td>(0.43%)</td>
</tr>
<tr>
<td>6</td>
<td>congestive heart failure, cramps cramps</td>
<td>(1.38%)</td>
</tr>
<tr>
<td>7</td>
<td>diarrhea</td>
<td>(0.96%)</td>
</tr>
<tr>
<td>8</td>
<td>fits</td>
<td>(0.32%)</td>
</tr>
<tr>
<td>9</td>
<td>heart attack Prednisone Allergic disorders, Anxiety, dizziness, depression,</td>
<td>(0.43%)</td>
</tr>
<tr>
<td>10</td>
<td>acid reflux skin conditions, insomnia, headache, nausea, depression</td>
<td>(2.97%)</td>
</tr>
<tr>
<td>11</td>
<td>Depression fever, headache, pharyngitis, cough</td>
<td>(1.12%)</td>
</tr>
<tr>
<td>12</td>
<td>Ditching difficulty with memory</td>
<td>(1.13%)</td>
</tr>
<tr>
<td>13</td>
<td>Topamax Seizures, migraine Anorexia, paresthesia (tingling), Tingling</td>
<td>(5.64%)</td>
</tr>
</tbody>
</table>

Figure 11. from root sider data to Chem ID
Figure 12. Root node - Chem Id to - Drug Name

Figure 13. with 2 drugs
Figure 14. with 2 drugs

Figure 15. 6 drugs
Figure 16. Network of drugs

Figure 17. Network of drugs
CONCLUSIONS

We have presented a novel Hidden Markov Model based text mining system that is capable of extracting adverse reactions of drugs based on the content available from online healthcare forums. We have shown that the information extracted from this system matches published information available on Drug Package Inserts. In addition we have also been able to identify some novel Adverse side-effect information that can act as early indicators for health authorities to help in their efforts towards Pharmacovigilance. The results are encouraging to pursue further enhancements to this approach.

REFERENCES


[22] Chen Y, Sampathkumar H, Luo B, Chen Xw: iLike: Bridging the semantic gap in vertical image


