International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 9 25 - 28

Identification of Association between Prescription Drugs and Side Effects by Analyzing Social Network Messages

K. Santhosh Kumar
Assistant professor, Department of Computer Science & Engineering
Annamalai University
Annamalai Nagar, India.
santhosh09539@gmail.com

P. Sudhakar
Assistant professor, Department of Computer Science & Engineering
Annamalai University
Annamalai Nagar, India.
kar.sudha@gmail.com

***Abstract*—** In this world of internet and social media all people have started discussing about their health information and treatment procedures in the health forums and social media like twitter. Researches are now being focused towards identifying hazardous effects of the prescription drugs and the treatment process through mining this information posted over the internet. Specifically, Twitter can be considered as an important source of information for the detection of such as Adverse Drug Reaction (ADR). The mining or analysis of Twitter messages is not easy because they are of short length, unstructured and almost in informal form. The twitter messages related to drugs prescribed for cardio vascular and diabetes were considered and collected to form the initial dataset. Later they are preprocessed to remove redundancy and improve the further classification process. A set of feature like Semantic, Z-Score, lexicon related features were extracted from the collected tweets to from the training dataset. Next the feature selection is performed using the Pointwise Mutual Information (PMI) approach. Finally, the selected feature set is utilized to train the Support Vector Machine (SVM). The SVM is validated with a test dataset and its performance was found satisfactory when linear function is used as the kernel. This model can be utilized further to identify the association between prescribed drugs and adverse effects from the Tweets and other messages of health forums.

Keywords- *Adverse Drug Reaction (ADR), Pointwise Mutual Information (PMI), Support Vector Machine (SVM), Twitter, Z-Score*

*****

1. INTRODUCTION

In the field of medicine, the therapeutic or adverse effect caused as a secondary along with or without intended effect is considered as the side effect. Often the term side effect is used to mention the adverse effect; sometimes it can be positive or beneficial. It is always the consequences of the usage of the prescribed drugs caused without intention. The side effect is the results of the prescription of the incorrect or unsuitable drugs or treatment procedure and is known as medical error. The side effects are created due to improper treatment and so they are referred as iatrogenic. The side effects will start when the dosage or treatment is started, increased or discontinued. The side effects may or may not cause complication of the disease or treatment procedure. The harmful effects of the drugs were indicated by abnormal condition like morbidity, mortality, change in body weight, modification in the level of enzyme secretion, function losses, or as pathological changes affecting prognosis. By the year 2013, because the side effects of the medical treatment and its side effects there were 142,000 deaths. Considering India, it has only less data and information regarding the side effects related to a prescribed drug. Also this information is considered very little when compared to the size of the pharmaceutical market of India. The World Health Organization (WHO) maintains a complete list of drugs and its associated side effects in database [1].

In 2015 India provided just only 2% of the 21Lakhs suspected adverse effects to the WHO's largest database, VigiBase [2] whereas China contributed around 8% to the database. At this juncture the analysis of the side effects associated with a particular drug is required and a complete database of information must be provided which can create awareness among the consumers and physicians. In this era of Internet, the consumers are reporting information related to the side effects and its review on the internet forums and social networks. These reports can be mined using machine learning technique to extract a list of drug and its side effects.

Even though the Indian government runs many monitoring program to access the side effects it still requires support from the drug manufacturers and other agencies to form a complete database. This work proposes a method to mine the messages posted in the world's largest social network twitter to identify association between the drugs and its side effects. It is believed that the social networking sites and online health forums contains an abundant of information regarding the adverse effects of the medicines and the treatment procedures. People have started posting quality reports on the internet which is similar to a report from a health professional. The recent research has showed that the patients report available over the internet have helped health department and drug manufacturers to identify the adverse effects which are undiscovered previously. Machine Learning algorithms are used in this work to classify the reviews and comments about a drug posted on the forums, either adverse or positive [3]. The Fig. 1 Presents the methodology adopted in this work for extraction of adverse side effects.

The analysis of the messages posted over the internet using Natural Language Processing algorithms and methods is very challenging as they are highly unstructured and informal in nature and the drug name mentioned in them may be misspelled. To mine the association between the drugs and its side effects from the social media postings initially the social media postings that reference to a prescription drugs must be identified. Even if the messages or postings are structured automatic identification of association between the side effects and the drugs is a complex task. This complexity is due to the fact that the relationships between drugs and their side effects, positive effects, indications are complex. Consider an example

tweet like if a patient posted that the drug makes me sleepy, it may be an adverse or positive effect. Also to form a training dataset that is used to train the statistical models used for classification need manual annotation which also increases the complexity. This paper is focused on mining tweets related to a set of commonly prescribed drugs for curing diabetes and cardio vascular diseases [4]. The proper identification of tweets including the misspelt drug names, sufficient pre-processing, correct manual annotation is needed to optimize the classification accuracy of the models trained using this data. The classification is performed using two class machine learning algorithm, Support Vector Machine (SVM).

2. DRUG LIST & TRAINING DATASET

The list of drugs that are focused must be prepared for further analysis. The list of drugs commonly prescribed for diabetes and cardio vascular diseases are mentioned in the list given by the Central Drugs Standard Control Organization of India is considered in this research [5]. Only two diseases are considered since they are commonly prevailing in all categories and segment of the people. As misspellings are forecasted in the tweets and messages posted in the twitter misspellings are generated using a phonetic misspelling filter. Only a portion of the misspellings are added to list.

Once the final drug list is prepared the tweets are accessed through the twitter API's available publicly. Using the API, it is possible to extract 1000 tweets per day related to the search topic. Nearly 5000 tweets were collected related to the search disease. Next the dataset is balanced to avoid the dominance of a particular drug whose tweets are more in number. On an average 100 records were prepared for a particular drug. All twitter datasets were stored in a repository. As another pre-processing step the stop words present in each of the tweets are removed. Normally when working with text mining applications the term stop words are discussed [6]. The stop words are nothing but common words used in English language. It is essential to remove the stop words from the tweets before features are extracted from it since it can help us to focus on the keywords.

For example, consider the tweet "After consuming Metformin kidney problems come for me". In this tweet the words like "after, come, for, and me" can be considered as stop words which are frequently occurring in English phrases and sentences. Next the collected tweets are to be annotated manually to form a training dataset for supervised learning. The tweets are labeled either as adverse effect or positive effects including neutral sayings. This is done after consultation with an experienced physician. The whole process of identification of association or relation between the drugs and its side effects is presented in the Fig. 2. The only difficult and time consuming task is the manual annotation of the tweets either as positive or negative effect.

3. FEATURE EXTRACTION

From the lexicons constructed from tweets taken four features like the number of positive words, number of negative words, the number of positive words divided by the negative words and the polarity of the final word are extracted [7]. Also the sum of positives scores and the negative scores are extracted from lexicons constructed automatically. As the term frequencies follow the multinomial distribution and the z-score can be considered as the standard form of the term frequency. The Z score is computed for each term in a class by finding its term relative frequency and the standard deviation of each term [8]. The Z- score is computed mathematically using the formula given below

$$Z_{score}(t_i) = \frac{tfr_{ij} - mean_i}{sd_i} \qquad (3.1)$$

The number of words having Z-score higher than the threshold is added to the feature vector along with the Z-score. The semantic feature representation of the text corpuses can provide some hidden information which is best suited to represent text data. The semantic representation of a text may bring some important hidden information, which may result in a better text representation and a better classification system. Each word in the tweets are mapped to their respective cluster in Brown, approximately 1000 features are added to the feature vector. Each feature in the vector represents the number of words in the tweet mapped to each cluster. Some of the other features used in this work are terms presence and frequency and these are n-grams words with their frequency counts. And finally the negative words present in the tweets if any are also added to the feature vector.

4. CLASSIFICATION OF THE TWEETS

Classification or analysis is the branch of Text Mining which deals with the process of extracting meaningful information from unstructured tweets. The classification of tweets posted related a drug in to either positive or adverse effects can be considered as sentiment classification which is also a binary classification problem. This is not equivalent to opinion mining in the sense that it may have numerous classes. In this work the support vector machine with linear kernel is used for classification. The linear kernel is feaster and easy to learn and also they can provide a good generalization accuracy. The other two non-linear kernels are capable of providing only small merits when compared to the linear kernels. The Sequential Minimal Optimization is used in our experiments to learn vector of feature weights. After learning the feature weights, the model now can classify new items using the new weights computed and the feature vector computed from the test tweets. The two parameters of the sigmoidal function that are capable of transforming the SVM output in to probabilities are also learned in our experiments.

The SVM classifiers perfectly suits for the classification of Text data as they are sparse in nature. Few of the features are irrelevant still they a tendency to be correlated with each other and can be categorized as linearly separable category. The SVM are capable of building a non-linear decision surface in the feature space by mapping instances of the data to a product space non linearly where the classes can be separated using a hyper plane.

5. FEATURE SELECTION METHODS

In the future selection process the task is to go for a considerable reduction in the size of the feature set without affecting the performance of the classifier. Feature selection can be lexicon based or statistical method based one. The

commonly used and the method used in this work is Pointwise Mutual Information (PMI). The PMI represents the association used in the statistics and information theory. By the PMI method the association between the unknown features set and the Positive or Negative term in the tweet are used to extract the sentiment of that corresponding feature. After computing the PMI score for each term, the feature vectors corresponding to words whose PMI is above a certain threshold are considered to training the classifier.

6. EXPERIMENTS & RESULTS

In this section, we present the experimental results of the binary classification process carried out using the Support Vector Machine (SVM). The SVM is a kernel machine suited for most of the text classification operation. The linear kernel utilized in this research also suitable when the size of the feature set is larger. This is because the mapping of data to a higher dimensional space will not increase the classification accuracy model. Here in our case both the number of records and the number of features are large when compared to a common image processing kind of application. Also training a SVM model with linear kernel is also faster when compared to other kernels also the number of parameters that are needed to be optimized is less in the case of linear kernel. With this background the linear kernel is well suited for text classification. The performance of the SVM classifier is not compared with other classifiers since for the binary classification it best suited. The tweets related to the interested disease are preprocessed and set of relevant features are extracted then the SVM is used to decide whether an adverse effect is mentioned in it or not. The classifier is trained in a supervised mode using the features extracted from the tweets annotated manually.

The proposed system achieves a rate of 75% in F1-score. This is comparable to other system proposed in various other researches. For a preliminary evaluation the constructed SVM model is tested with the twitter stream. Because of the systems' low precision nearly 30 of the tweets were classified as containing adverse effect. This is too high when compared to the number of actual tweets containing adverse effects of the drugs. For this reason, the tweets were preprocessed to contain tweets related to certain disease. When the model is applied on the reduced set it found 800 tweets containing ADR. It is expected that nearly 75% of tweets contains information related to drug adverse effects. Using cross validation, the evaluation parameters such as precision, recall, f-measure and accuracy are estimated to validate the performance of the SVM classifier in text classification. The results are presented in Table 6.1. The analysis is made with dataset which is heavily imbalanced i.e. it contains 70% of the records text describing the adverse effects of drugs. The results indicate that accuracy of the classifier is high when the proportion of the majority class is also high. The analysis of the results revealed that if the number of instances of the majority class increases in the training dataset then it automatically increases the accuracy of the majority class. Also the overall accuracy depends on the majority class; if the accuracy of the majority class is increased then the overall accuracy is also increased.

Table 6.1 Performance of the SVM classifier

Records with no ADR			Records with ADR			
Precision	Recall	F-Measure	Precision	Recall	F-Measure	Accuracy
Balanced Dataset						
78.5	67.3	72.6	61.8	74.6	67.8	70.8
Dataset with 60% ADR						
84.3	73.5	78.2	44.95	65.1	53.1	71.9
Dataset with 70% ADR						
88.9	78.9	83.8	43.6	64.7	52.5	75.9

Experiments are also performed with the balanced dataset. The dataset is balanced using down sampling of the records with the adverse drug effects. As the proportion of the records with ADR increases the accuracy is increased. This experiment does not focus on finding an optimal proportion of the classes or cut-off limit percentage of majority class in the dataset. The Fig. 3 is the graphical representation of the experimental results.

7. CONCLUSION

In this work an annotated twitter corpus is constructed for identification of ADR mentioned covering a broad set of drugs related to the cardio vascular and diabetes diseases. Only 65% of the drugs mentioned in the list were present in the tweets posted by the public people. The data set is balanced using down sampling to avoid the dominance of a particular drug in the dataset. The pre-processing refines the dataset before features are extracted from them. This work is designed after reviewing the social media posts and messages over a period of time regarding the adverse effects of the drugs prescribed by the physicians. But in the public domain there is no annotated dataset describing the adverse effects. To strengthen the pharmacovigilance program of the Indian government it is essential to build an annotated dataset and should be made available to public research. In this work a base work has been implemented to demonstrate the usage of public messages posted in the social media for the identification of adverse effects. It is shown that using the supervised machine learning approaches the task of identifying whether a tweet contains the adverse effect mentioning in it or not. Further this research work can be extended using other Natural Language processing algorithms and multi class classification algorithms for the detection of the adverse effects in the social media messages.

References

[1] https://www.drugwatch.com/side-effects/
[2] Lindquist M. VigiBase, the WHO global ICSR database system: basic facts. Drug Inf J 2008; 42: 409-19.
[3] Gurulingappa, Harsha, Abdul Mateen-Rajpu, and Luca Toldo. "Extraction of potential adverse drug events from medical case reports." *Journal of biomedical semantics* 3.1 (2012).
[4] Ginn, Rachel, et al. "Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark." *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*. 2014.

[5] McGettigan, Patricia, et al. "Use of fixed dose combination (FDC) drugs in India: central regulatory approval and sales of FDCs containing non-steroidal anti-inflammatory drugs (NSAIDs), metformin, or psychotropic drugs." *PLoS Med* 12.5 (2015): e1001826.

[6] Munková, Daša, Michal Munk, and Martin Vozár. "Influence of stop-words removal on sequence patterns identification within comparable corpora." *ICT Innovations 2013*. Springer International Publishing, 2014. 67-76.

[7] Collins, Riley, et al. "SWAT-CMW: Classification of Twitter Emotional Polarity using a Multiple-Classifier Decision Schema and Enhanced Emotion Tagging." *SemEval-2015* (2015): 669.

[8] Hamdan, Hussam, Patrice Bellot, and Frederic Béchet. "The impact of z score on twitter sentiment analysis." *In Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*. 2014.

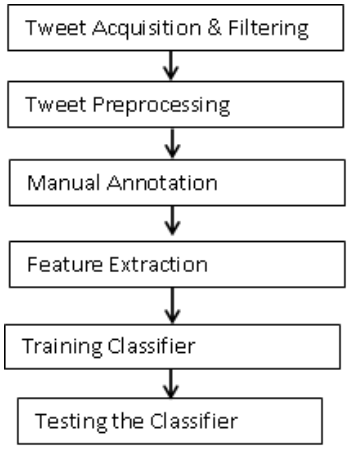

Fig 1. Pictorial representation of the proposed methodology

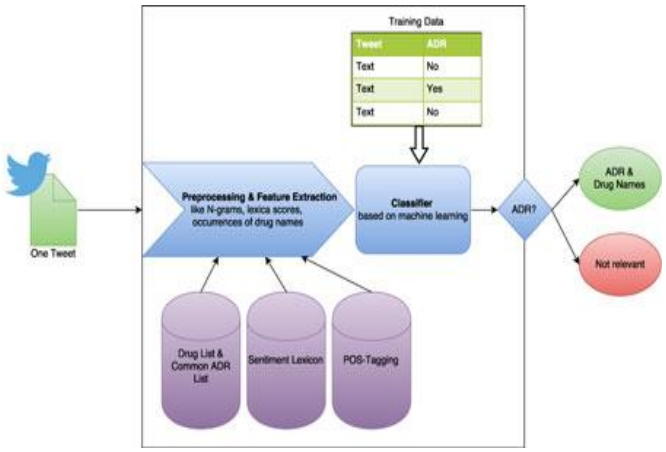

Fig.2 Process flow of identification of association

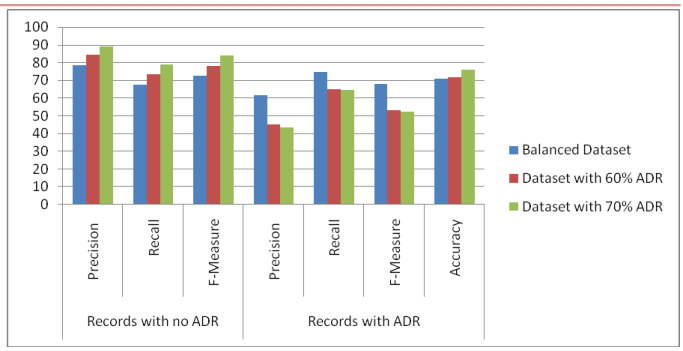

Fig 3 Analysis of classifier peromance