COMP90049 Project 2 Report: Identifying Tweets with Adverse Drug Reactions

Shuai Wang

1. Introduction

The aim of the project is to assess the effectiveness of Naïve Bayes machine learning classification method on the problem of determining whether a tweet contains an Adverse Drug Reaction(ADR), and to express the knowledge that I have gained.

2. Data Set

The data set used in the project includes three parts: training, developing, and testing data. [1] Training data is used by WEKA to train a Naïve Bayes machine learning model. Developing data is used to evaluate the model with evaluation metrics. The model will do prediction on testing data. Each data part contains a *.txt file and a *.arff file.

The *.txt file collects many raw tweets extracted from Twitter with following format:

tweet-id TAB class TAB text-of-tweet

"tweet-id" is an unique identifier to a tweet.

"class" is either 'Y' or 'N', respectively indicating whether a tweet contains an ADR or not.

"text-of-tweet" is the content of a tweet.

The *.arff file contains all vector representations of tweets above. The file is suitable for use with WEKA. There are 92 best terms that are indicative to an ADR. We have counted their frequency in each tweet.

One @RELATION line gives the name of the data set.

One @ATTRIBUTE line for each attribute(term), including the name of the attribute and the type of value. Most of them are "NUMERIC". A nominal attribute like "class" can take any of values in braces {}.

One @DATA line indicates that the following lines are instances.

3. Naive Bayes

In this project, Naïve Bayes method [2] is used to classify a tweet $X = < x_1, x_2, \dots, x_{93} >$ according to one of classes $c_j \in \{Y, N\}$

$$c = \mathrm{argmax}_{c_j \in \{Y,N\}} P(c_j|X) = \mathrm{argmax}_{c_j \in \{Y,N\}} \frac{P(X|c_j)P(c_j)}{P(X)}$$

$$P(X|c_j) = P(< x_1, x_2, \dots, x_{93} > |c_j) = P(x_1|c_j) \cdot P(x_2|c_j) \cdot \dots \cdot P(x_{93}|c_j) = \prod_{i=1}^{93} P(x_i|c_j)$$

With two equations above, we can classify a tweet into the class c that has highest probability of $P(c_j|X)$.

4. New Attribute – "account"

A new attribute that may be helpful to determine whether a tweet has an ADR is "account". Twitter provides a special functionality for users, which is '@'. If one hopes someone can see his/her tweet, he/she can @ them. Hence, if a user wrote a tweet and @ someone who is a doctor or a medical specialist, the tweet is more possible to contain an ADR.

In the training data set, there are 3,166 raw tweets in which 1,114 tweets that @ someone else, and 152 such tweets contain an ADR. On the other hand, there are 2052 tweets that do not @ anyone, in which 82 such tweets do contain ADR. Hence, we can calculate the probability that a tweet @ someone and contains ADR, P(@, ADR)=13.6%. We can also compute the probability that a tweet does not @ anyone and contains ADR, P(^@, ADR)=3.9%. By comparing the two probabilities, we conclude that if a tweet @ someone, it will be more possible to contain an ADR.

Furthermore, if a tweet not only @ someone but also the one is a doctor or a medical specialist, the tweet has a high chance to include an ADR. There is a "train_accounts.txt" file in which it contains 201 medical specialist twitter accounts generated from training data set. If a tweet contains one account in this file, its value of the attribute "account" takes 'Y', otherwise takes 'N'.

5. Result

Firstly, use Naïve Bayes model to evaluate "dev" data with WEKA, and just represent a tweet over 92 terms. There are three tables as results of evaluation.

The Accuracy and Error rates are:

Indicators	Instances	Values
Accuracy	884	82.1561 %
Error	192	17.8439 %

Table 1 accuracy and Error rates

The Confusion Matrix:

Instances	(predicted) N	(predicted) Y
(Actual) N	829	133
(Actual) Y	59	55

Table 2 Confusion Matrix

Some tweets containing an "@account" with wrong predictions:

Tweet-id	line	actual	predicted	result	prediction
330249629247610880	53	Y	N	×	0.893
341257005232697344	374	Y	N	×	0.661
348907962372349952	852	Y	N	×	0.712
332103366001963010	85	N	Y	×	0.617
333519110946291712	169	N	Y	×	0.613
340654242270425089	338	N	Y	×	0.544
345787515506130944	702	N	Y	×	0.575
349042809526947840	862	N	Y	×	0.607
354302147036319744	1070	N	Y	×	0.598
354352302028627968	1073	N	Y	×	0.518

Table 3 Tweets with wrong predictions

Now, add the new attribute "account" into vector representation. Use Naïve Bayes model evaluate again. There are three new tables:

Accuracy and Error rate:

Indicators	Instances	Values
Accuracy	891	82.8067 %
Error	185	17.1933 %

Table 4 Accuracy and Error rates

The Confusion Matrix:

Instances	N	Y
N	843	119
Y	66	48

Table 5 Confusion Matrix

For tweets above got corrected:

Tweet-id	Line	actual	predicted	result	prediction	account
330249629247610880	53	Y	Y	√	0.993	@ecrjones
341257005232697344	374	Y	Y	√	0.998	@LithiumLibGirl
348907962372349952	852	Y	Y	√	0.998	@LithiumLibGirl
332103366001963010	85	N	N	√	0.511	@elladeruiter
333519110946291712	169	N	N	√	0.516	@tete_floue
340654242270425089	338	N	N	√	0.586	@MJIsWithMe
345787515506130944	702	N	N	√	0.555	@Sarah_G_Nelson
349042809526947840	862	N	N	√	0.522	@NerdyLori
354302147036319744	1070	N	N	√	0.532	@anwen
354352302028627968	1073	N	N	√	0.611	@Rachs_charlie

Table 6 Tweets with correct prediction

6. Analysis

Firstly, the accuracy rate increased from 82.1% to 82.8%. This means the new attribute "account" is helpful to determine whether a tweet contains an ADR. From examples above, some tweets with wrong predictions have been corrected as "@account" indicates the tweet has more probability to contain ADR.

Secondly, by comparing two confusion metrics, we found the amount of NN tweets (actual class is N and predicted class is N) got increased. This is because some tweets that do not contain an ADR with prediction "Y" got corrected. On the other hand, the amount of YY tweets (actual class is Y and predicted class is Y) got decreased. This is because the model with new attribute made wrong prediction. Some tweets contain ADR but they might not have any twitter account, which leads their values of "account" take "N". This affects the prediction.

Thirdly, by observing those tweets containing ADR, I found some of them do not contain any "@account", and some of them do contain "@account" but those twitter accounts are not a doctor or medical specialist. Thus, the "account" attribute like other attributes might exist in a tweet or not. It depends on how the user describe his/her reaction after taking some medicine. All accounts in the "train_accounts.txt" file are derived from raw tweets in "train.txt". Some of them might not be relative to medical staffs. This is one drawback of the attribute. If there is an account data set that contains plenty of medical staff Twitter accounts, the model will predict better.

Fourthly, some "@account" (Twitter accounts) in tweets do not exist now. If the Twitter account does not exist in real world, one tweet @ this account. It will not provide any indicative information for us. Hence, we should do pre-process to eliminate those meaningless accounts in tweets.

7. Conclusion

In the project – Identifying Tweets with Adverse Drug Reactions. The "account" is introduced as an additional attribute. The Naïve Bayes machine learning algorithm is used to determine whether a tweet contains an ADR or not. Some Java programs have been implemented to help calculate the value of "account" for a tweet. WEKA has been used to train Naïve Bayes model, evaluate the performance and do prediction. The "account" attribute is somewhat helpful. However, it has some drawbacks.

8. Reference

[1] Abeed Sarker and Graciela Gonzalez. (2015) *Portable automatic text classification for adverse drug reaction detection via multi-corpus training*. Journal of Biomedical Informatics, 53: 196-207.

[2] Naïve Bayes Classifier. Retrieved from https://en.wikipedia.org/wiki/Naive_Bayes_classifier.