Identifying Tweets with Adverse Drug Reactions

1.Introduction:

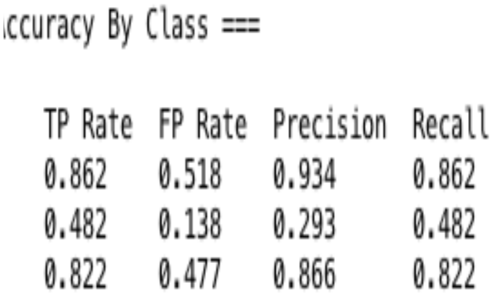
Nowadays, the information provided by social media like Twitter contains incredible amount of information that may potentially benefits our life. Many research studies with in recent years have shown that Twitter is a useful resource for ADR detection. “Tweets can be used to supplement existing electronic ADR monitoring systems by providing real-world, real- time clinical narratives from users posted in the public domain”.[1] One of the recent researches develop the technic to distinguish indications of personal medication intake in some social media. This paper also shows a method, which could be used to train the machine-learning model to determine whether a tweet that mentions a medication indicates that the individual posting has taken that medication[2]. In this paper, we will focus on analysis the outcome generated by Naïve Bayes Model and improve the performance of that model.

The dataset we are using is from Twitter, and is an altered form of a dataset from the DIEGO Lab[3]

2.1 Naïve Bayes Introduction:

While constructing classifiers， Naive Bayes is a simple and wide used method. The NB models, which represented as vectors of attributes value, assign class labels to each instance. In this case, the class label is “N” and “Y”, where “N” means the poster has ADR and “Y” implies non-ADR tweeter user. One major character of Naïve Bayes classifier is that it assumes that all feature are mutually independent. Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector representing some n features (independent variables), it assigns to this instance probabilities.[4]

2.2 Naïve Bayes result Analysis:



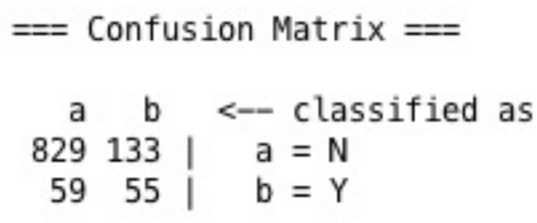
```
ccuracy By Class ===

  TP Rate  FP Rate  Precision  Recall
  0.862    0.518    0.934      0.862
  0.482    0.138    0.293      0.482
  0.822    0.477    0.866      0.822
```

[1] Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark,. Rachel Ginn1, Pranoti Pimpalkhute1, Azadeh Nikfarjam1,, Arizona State University

[2] Detecting Personal Medication Intake in Twitter: An Annotated Corpus and Baseline Classification System,
Ari Z. Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O’Connor, Graciela Gonzalez, Brandeis University

[3] 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.

[4] https://en.wikipedia.org/wiki/Naive_Bayes_classifier



```
=== Confusion Matrix ===

   a   b   <-- classified as
 829 133 |   a = N
  59  55 |   b = Y
```

The TP Rate measures the proportion of tweets that contain no ADR that are correctly identified. In this case, just as the confusion matrix shows, the total number of N class tweets is 962 and 829 of them are assigned to the right class, which implies that 86.2% of the tweets from N class are identified correctly. For Y class, our model detects only 55 out of 114 tweet with ADR information from the target data set and thus the TP rate for Y class is 0.482.

The FP rate measures the proportion of all tweets that contains ADR information that still yield positive test outcomes. Almost half of them were detected by the Naïve Bayes method. Since the FP rate of N class is much higher than that of Y class and a higher precision means less false positives, the precision of N class is much 0.934, which indicates that the NB model has very high accuracy to predict the N class members. However, since only half of the Y class members are predicted correctly, the precision of Y class are pretty low, which implies that the prediction for Y class tweets is not that reliable.

One of the possible reason that the accuracy of Y class is lower than that of N class is that, when we calculate the outcome probabilities via Naïve Bayes, we implicitly assume that all attributes are mutually independent and therefore we could calculate the outcome simply by multiply the conditional probabilities.

However, in our data set, most of the attributes are not independent with each other. Since our attributes are nominal or categorical, we could perform a logistic regression to get the p-value to exam the dependence relation between those variables. Based on the results, we can't indicate that the attributes are independent or the relation between those features will not affect the final result. Some of the attributes are even closely related. For example, bed and sleeps are always referring to the same behavior. Gain and gained are express the same meaning. All those related attributes are treated as separate and independent to each other, which leads to bias.

Since we are using binary expression for most of time to identify the attributes, the relation between those attributes could affect the Y class more since Y class is more sensitive to the depend relations between the feature. The number of tweets in N group is much more than that of Y group. Since the sample size is much larger, the training data contains more accurate information for the prediction and the implicate relation may not affect the final result. However for Y class, since the sample size is pretty small, the outcome provided by Naïve Bayes model will be more sensitive to the factor of dependent relationship.

3.1 New attributes Introduction:

For all 93 features provided, some of them could affect the final result more than others. If we could find the feature, which contributes more to the outcome and refine the attributes based on the corresponding data, the overall performance may improve. Thus, I use the Infogain feature selection to ranking the each attribute by measuring the information gained with respect to the class. Based on the result, words like me, makes, feel, I, lozenge, sick, Prozac, it, making, pain and etc. are contributes more to the model than words like pic, sense, appetite. If we look up the corresponding sentence in the original tweet file, this situation will become more clearly.

First, since people usually talks about the adverse effects in post-approval drugs constitute about themselves rather than someone else. Thus, the subject or the major of the tweets are always the poster themselves. For the word like "I", "my", "me " are ranking very high in our selector. For instance: "*I think my tablets have made me gain weight. Anyone on fluoxetine/prozac? #replytweet"*, *"Thought of work is overwhelming me so much I feel like crying but can't because olanzapine has me trapped in a zombie state #strugglewithin", "@pseudodeviant i clock in at 16 stone (added 1 stone since quetiapine) and according to that BMI thing I'm obese, which I'm clearly not."* and *"But I don't want to take an ambien or trazodone or anything because i dint wanna restart my dependence upon them",* the first four sentence from the train.txt dataset are all talk about the poster themselves with the word "i", "my" and "me".

Second, most of time, the patient are passive accepters of the ADR, the verb like make, gain, feel also play an important role in building the prediction model.

Thirdly, the corresponding drugs, like lozenge, Prozac, olanzapine, rivaroxaban, chemical, used by people who surfer from ADR is also highly related with the outcome the model. For example: "Swollen feet thx to olanzapine", "I'm also on Clonazepam, makes me sleepy too", "first place was due to the drug olanzapine" and "part of withdrawal symptoms from Venlafaxine"

3.1.1 Method 1:

Since some of the attributes are expressing the same meaning or are highly related, in order to increase the overall performance, we can add a new attributes by check the synonyms. For example, "I, my, me, am" are expressing the same meaning, thus we can build a set of synonyms and if the sentence contains any of the word in the set, we assign value 1,, otherwise, we assign 0.

3.1.2 Method 2:

In order to reduce the dependent relationship between different features, we could also add new attributes that summarize a category of attributes. For instance, Cymbalta, glassncision, lamotrigine, lozenge, olanzapine, nicotine, Prozac and etc. could be represented as medicine. If we use medicine to substitute those words, the side effects caused by the implicated relationship between those drugs with same purpose will be reduced.

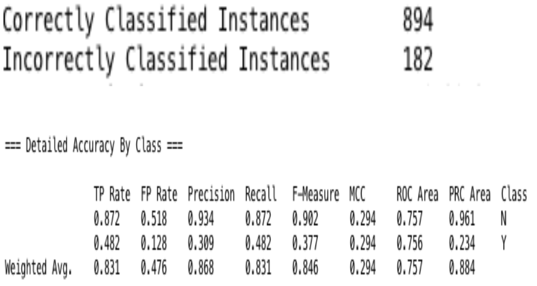
Also, based on the dataset, most of time, people with ADR always use the sentence structure like some medicine make me feel something or some feeling caused by some medicine to express their idea. Thus we could also add this feature to the model to detect the sentence construction to predict the outcome.

3.1.3 Method 3:

When people post tweets, sometimes they

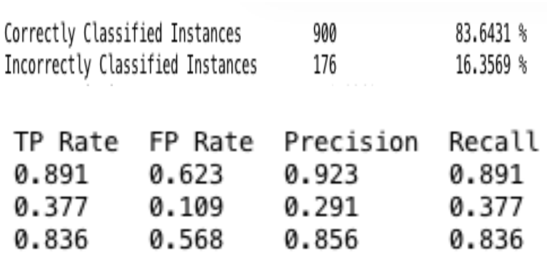
tend to use some repeating letters to exaggerate their feeling. Word like “sooooo”, “gooooo”, “saaaad” may contributes to the final outcome in some extend. However, in this case, the result shows that this method is useless. When people talk about heavy topic like ADR, they are prefer to express the real feeling rather than exaggerated expressions.

3.2 New attributes result Analysis:



```
Correctly Classified Instances         894
Incorrectly Classified Instances       182


=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.872    0.518    0.934      0.872    0.902      0.294    0.757     0.961     N
               0.482    0.128    0.309      0.482    0.377      0.294    0.756     0.234     Y
Weighted Avg.  0.831    0.476    0.868      0.831    0.846      0.294    0.757     0.884
```

From the evaluation matrix and summary above, after we add the new contribution subject and delete the repeated synonyms attributes (I, am, my, me), the correctly classified instance is improved slightly as well as the precision for Y is also increased.



```
Correctly Classified Instances         900               83.6431 %
Incorrectly Classified Instances       176               16.3569 %

 TP Rate   FP Rate   Precision   Recall
 0.891     0.623     0.923       0.891
 0.377     0.109     0.291       0.377
 0.836     0.568     0.856       0.836
```

From the evaluation matrix and summary above, after we add the new attribute subject and a new attribute medicine (Category of terms: medicine=chemical, Cymbalta, glassncision, lamotrigine, lozenge, olanzapine, nicotine, Prozac, quetiapine, rivaroxaban) and delete the repeated synonyms attributes and the attributes belongs to the medicine category, as well as we also checked the sentence structure (“medicine +make/made/making/ sb” and “caused by + some medicine”) , the correctly classified instance is improved 0.8%.

4. Conclusion:

After we reduce the dependent relationship between different attributes and by using synonyms set, category set and sentence structure checking to construct the new attributes, the performance of the model is improved.

Reference

[1] Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark,. Rachel Ginn1, Pranoti Pimpalkhute1, Azadeh Nikfarjam1,, Arizona State University, Available at: http://www.aclweb.org/anthology/W17-2316

[2] Detecting Personal Medication Intake in Twitter: An Annotated Corpus and Baseline Classification System, Ari Z. Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O’Connor, Graciela Gonzalez, Brandeis University, Available at : http://aclweb.org/anthology/P16-3003

[3] 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.

[4]https://en.wikipedia.org/wiki/Naive_Bayes_classifier