COMP90049 Knowledge Technologies
Project 2: Identifying Tweets with Adverse Drug Reactions

1 Introduction

The goal of this report is to find the method of predicting whether a Tweet contains ADR (Adverse Drug Reactions). We will use 92 word-unigram attributes to build a machine learning model, discuss the selection of several new non-word-unigram attributes for the model and analyze the performance of the model.

2 Data Set

The data we use for model building and testing in this report is from Twitter, and the form is modified by Sarker and Gonzalez (2015). The data set consists of three parts: train, dev and test. Each part contains an arff file with the value of 92 attributes and a text file with tweet id, tweet contents and a class label showing whether the tweet contains ADR. All the tweets in the data set are associated with drugs. In this report, we use train data set to train the model, and use dev data set to evaluate the model.

3 Evaluation Metrics

There are four types of result for our problem:

- TP: True Positive, Positive instances classified as Positive
- FN: False Negative, Positive instances classified as Negative
- FP: False Positive, Negative instances classified as Positive
- TN: True Negative, Negative instances classified as Negative

In this report, we use precision and recall proposed by Fraser and Marcu (2007) for evaluation.

- Precision: proportion of correct result among all positive predictions, which is $\frac{TP}{TP+FP}$.
- Recall: proportion of positive tokens with correct prediction, which is $\frac{TN}{TP+FN}$.

The performance of the model will be evaluated based on the precision and recall for "Y" class, "N" class, and weighted average.

4 Original Model

In this report, we build machine learning models by a machine learning tool with GUI, Weka. The machine learning model is built by Naive Bayes using the given 92 attributes which are term counts in tweets. Instances with label Y are tweets containing ADR, while instances with label N are tweets not containing ADR.

Naive Bayes predicts the class of data according to the joint probability distribution of attributes and is based on the assumption that all attributes are independent(A.McCallum and K.Nigam (1998)). An instance X with attributes value $< x_1, x_2, ..., x_n >$ is classified to class c where

$$c = argmax_{c_j \in C} P(c_j) \prod_i P(x_i|c_j)$$

The result of Naive Bayes model is shown in Table 1.

	Class N	Class Y	Average
Precision	93.3%	28.8%	86.5%
Recall	85.9%	48.2%	81.9%

Table 1: Result of Original Naive Bayes Model

We can see from the result that precision and recall of class N is much higher than that of class Y, which means the classifier is more likely to categorize data as class N. This result may be because there are much more non-ADR tweets than ADR tweets in the data set and the model has a more accurate probability distribution of non-ADR related attributes.

5 New Attributes Analysis

In this section, we will discuss several new non-word-unigram attributes that may improve the performance of the original method. After observing the pattern of tweets containing ADR or not, I came up with idea of new features in the following three aspects.

5.1 Personal Feeling

The most evident characteristic of tweets containing ADR is that they include users' feelings or reactions to the drug. There are several typical ways to express users' feelings in a sentence in the data set:

- using first-person pronouns
  eg. I feel like crying but can't because olanzapine has me trapped in a zombie state
- make with object and complement
  eg. Morning fluoxetine burp makes me want to ralph.
- using drug name or it as nominal subject
  eg. I keep telling myself 8 am, but I think the Trazodone fucks it up.It knocks me right out and won't let me wake up.
- drug is the direct object of subject
  eg. Taking seroquel is like swallowing a sleeping pill when u wake up then trying to function normally

In this report, we will try to add new attributes based on first two sentence patterns.

5.1.1 First-person pronouns

To find the first sentence pattern in tweets, we can add a new attribute which indicates if a tweet contains first-person pronouns. Apart from finding first-person pronouns directly, I use Stanford CoreNLP Natural Language Processing toolkit (Manning and McClosky (2014)) to analyze if any nominal subject of the sentence is "I", which will identify "I" even in contractions such as "I'm" and "I've". The performance of model after adding this new attribute and removing attribute "am", "I", "my" and "me" is shown in Table 2.

	Class N	Class Y	Average
Precision	93.4%	31.4%	87.7%
Recall	87.7%	47.7%	83.5%

Table 2: Result of Naive Bayes Model With First-person Attribute)

We can see from the result that the recall of class Y and precision of class N improve while the recall of class N drops slightly. In other words, the new model can identify much more N class instances as N class than before.

The original model uses count of "am", "I", "my" and "me" as attributes, which may lead to some defects in classification. Firstly, Naive Bayes makes predictions based on the conditional probability of attributes. However, the number of non-zero first-person pronouns count is not related to whether the tweet contains ADR in general, which may lead to the overfitting of instances with high first-person pronoun quantity. Secondly, the appearance of some attributes (such as am and I) is correlated, but there is no attribute in the original model indicates the relationship between them. Also, it cannot identify "I" in contractions and "am" used as the period which will cause the inaccuracy of counts.

5.1.2 Make

The Original model has three attributes related to the second sentence pattern: made, makes and making. However, these attributes only consider term counts but not the object after them, which may cause the decline of model performance. We can improve it by identifying the appearance of these words followed by "me", "you" or objects including "my". Also, we can consider adding a new attribute "make". The performance of model after improving make attributes is shown in Table 3.

	Class N	Class Y	Average
Precision	93.6%	30.0%	86.8%
Recall	86.2%	50.0%	82.3%

Table 3: Result of Naive Bayes Model With Make Attributes Improved

We can see from the result that the performance is a bit better than the original model. This improvement is due to the reason that among all the usage of "make", it is more likely to be used for expressing personal feelings in tweets containing ADR than tweets not. Thus for all "make" related attributes, the gap of mean value between class Y and class N is larger than the previous model and is better for identifying the characteristic of each class. These improved attributes let the model work better in predicting sentences containing the "make" and conforming to the sentence pattern as class Y,

and sentences containing "make" but not conforming to the sentence pattern as class N.

5.2 Sentiment

Most tweets containing ADR are negative in sentiment because the user is unsatisfied with the drug effect while tweets without ADR may be more positive in average. The sentiment of the tweet can be expressed by emoticon and verbal contents.

5.2.1 Emoticon

Emoticons can imply the sentiment of the tweet. For example, emoticons such as ":)", ";-)", ":-D", ";-P" shows this tweet is positive while emotions such as ":\" or ":(" suggest this tweet is negative.The performance of model after adding new attributes is shown in Table 4.

	Class N	Class Y	Average
Precision	93.3%	28.9%	86.5%
Recall	86.0%	48.2%	82.0%

Table 4: Result of Naive Bayes Model After Adding Emoticon Attributes

As shown in Table 5, attribute positive emoticon has a significant mean value difference between two classes, but the performance only improves slightly. One reason for that is emoticons rarely appear in our data set and can make little difference to our original model. But it is good to see that the prediction value of tweets containing positive emoticons classified as class Y is lower than the previous model, which means the model is less confident about this prediction.

	class N	class Y
Mean	0.0308	0.008
Std. Dev.	0.0122	0.0268

Table 5: Attribute Positive Emoticon

5.2.2 Verbal Contents

The sentiment of verbal contents is analyzed using Stanford CoreNLP toolkit (Manning and McClosky (2014)). I represent the sentiment of the verbal text of tweets in three attributes: whether a tweet contains a positive sentence (positive), whether a tweet includes a negative sentence (negative) and the overall sentiment score. The results are shown in Table 6 and Table 7.

	Class N	Class Y	Average
Precision	93.1%	27.7%	86.2%
Recall	85.7%	46.5%	81.5%

Table 6: Result of Naive Bayes Model With Negative and Positive Attributes

	Class N	Class Y	Average
Precision	93.2%	28.0%	86.3%
Recall	85.6%	47.4%	81.5%

Table 7: Result of Naive Bayes Model with Sentiment-score Attribute

Both precision and recall drop for these two methods of indicating verbal contents sentiment. Our result shows that tweets with ADR are more likely to contain a negative sentence as expected but the probability of containing a positive sentence is similar. Also, there is no evident difference between the sentiment score of two classes, with mean value 0.5736 and 0.4198 respectively. The flaw in the sentiment-analyze algorithm is one reason for that. For example, "lithium makes my hand shake." is identified as a positive sentence. Apart from that, non-ADR tweets also associate with drugs, which means they are not positive in general. In this case, other terms in existing attributes are more correlated with ADR label and adding sentiment attributes may cause the drop of performance.

5.3 Hyperlink and Price

Hyperlink and price are much more likely to occur in non-ADR tweets, which means they are negatively correlated with ADR. Thus, the presence of the hyperlink and "$" in the tweet can be considered as new attributes. The result is shown in Table 8.

	Class N	Class Y	Average
Precision	93.3%	28.6%	86.5%
Recall	85.8%	48.2%	81.8%

Table 8: Result of Naive Bayes Model With Link and Price Attributes

Similar to emoticons, these two attributes have little influence to the original model because of the low occurrence in data set.

6 Conclusions

In this report, we build machine models using Naive Bayes to identify ADR in tweets. The original model is based on 92 word count attributes. Attributes related to personal feelings (first-person pronouns and make) can improve the performance of model. Attribute emoticon, link, and price have limited impact on the model. Attribute about sentiment in verbal contents can reduce the performance.

References

A.McCallum and K.Nigam. 1998. A comparison of event models for nave bayes text classification. *A Comparison of Event Models for nave Bayes Text Classification, In AAAI-98 Workshop on Learning for Text.*

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, pages 33(3):293–303.

Mihai Surdeanu John Bauer Jenny Finkel Steven J. Bethard Manning, Christopher D. and David McClosky. 2014. The stanford corenlp natural language processing toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

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