A Text Mining Pipeline for Classifying Healthcare Forum Posts

Oguzhan Gencoglu

Department of Signal Processing, Tampere University of Technology, Finland

Introduction

Patients and caregivers often share their health-related concerns in community forums and discussion boards. One of the important tasks towards automated question answering is to identify the topic of questions being asked by the patients and their caregivers. Here, we propose a text mining pipeline for classifying healthcare forum posts/messages/questions.

Data

The healthcare forum messages data consist of training and test sets, with 8000 and 3000 observations, respectively.

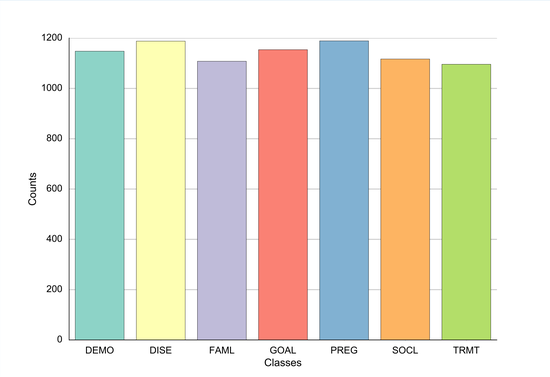

Figure 1: Distribution of 7 question classes in the training set.

The possible categories for a given message are *Demographic, Disease, Treatment, Goal-oriented, Pregnancy, Family Support* and *Socializing*.

Remark

All the decisions including *preprocessing steps*, *classifier choice* and *classifier parameters* (loss function, regularization coefficient etc.) have been chosen by 10-fold cross-validation and grid search over the training set.

Methodology

- *Preprocessing* -

For each observation in both training and test set:

1. *Title and the message are merged with a whitespace in between.*
2. *All text is converted to lowercase letters/characters.*
3. *A whitespace is added after each '.' or ',' unless it is already there.*
4. *Any number of consecutive whitespace characters are transformed into a single whitespace character.*
5. *All non-ASCII characters are removed.*
6. *All urls are removed.*
7. *Word stemming is applied on each word with the help of WordNet lexical database [1].*

- *Feature Extraction* -

The preprocessing step is followed by the feature extraction step for each observation in both training and test datasets. *Term frequency - inverse document frequency* (tf-idf) features are extracted from each observation [2]. English stopwords and an n-gram range of 1 to 3 (inclusive) are used for the feature extractor.

- *Classification* -

After the feature extraction, an *online passive-aggressive classifier* is trained on the training set [3]. The classifier loss to be minimized is set to be *hinge-loss*. In addition, the data is not assumed to be centered for the training procedure and a regularization coefficient of 0.3 has been set for the algorithm.

Remark

One advantage of the proposed pipeline is that when new data is available, the classifier can be updated instead of starting the training form the beginning.

Results

The proposed machine learning pipeline is trained with 8000 observations. The confusion matrix for the local 10-fold cross-validation can be examined in Figure 2.

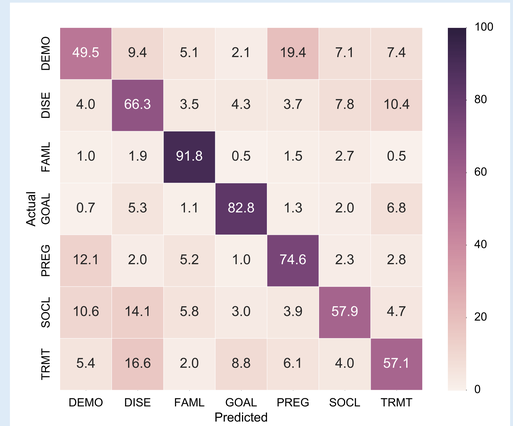

Figure 2: Confusion matrix of the local 10-fold cross-validation.

The accuracy on the test set of 3000 observations are presented in Table 1.

Table 1: Accuracies of local 10-fold cross-validation and test set.

	Accuracy
Local 10-fold CV	68.61%
Test Set	**67.47%**

References

[1] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.
[2] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003.
[3] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551-585, 2006.

oguzhan.gencoglu@tut.fi