CS59000
Homework 2

Penghao Wang

April 1, 2016

1 Introduction

The aim of this project is to classify stance of posts about different topics. There are three model implemented in this project. First is the model based on the bag of word model. The second local post-level model is based on the work of Anand et al.(2011). The third global model is based on the work of Hasan and Ng.(2013) The project used SVM and CRF algorithms from the python library to train and test the model.

2 Feature Set

In the project, two kinds of features sets are implemented. The first kind is the local post-level feature sets. These features set come from Cats Rule and Dogs Drool!: Classifying Stance in Online Debate (Anand et-al, 2011). The other kind of feature set is aim to frame the project as a global decision problem. There are two constraints are implemented in the project.

2.1 Local post-level

The feature sets are basically the same as the features in the Anand et al.'s approach (2011).

1. Bag of Unigram words. In the paper, the bigram and unigram model are implement as baseline. I only implement the unigram word list as the baseline.
2. The post length.
3. The word per sentence.(WPS)
4. The percent of the words that are longer than 6 letters.
5. The percentage of the positive or negative emotion words. In this part, the package form textblob is used. It helped me to calculate the percentage.

6. Opinion dependencies: I add the positive or negative score of entire post as the features.
7. The percent of the words a s the pronominal.
8. The bag of initial unigram words. It is reffered as cue word in the paper. It is used to capture the useage of cue words to mark responses of particular type, such as oh really, so, and well(Anand et al., 2011)
9. Repeated Punctuation. These repeated sequential use of particular types of punctuation are such as !! and ??.(Anand et al., 2011)
10. Post information as the rebuttal in the meta file.
11. The context features. Some features come from the parent post except the unigram.

These local features are implemented in the project. I will discuss about the accuracy in the section 3. The classifier is SVM model provided by sklearn package in Python.

2.2 The global model

There are some problems of the local model. For instance, there are more positive(stance is +1) posts than negative post. That means has a prediction bias of positive posts. The global model could solve this problem. In the global model, the series of articles are trained. In that case, it will improve the performance. There are two constrains are implemented in the author constraints and the user-interaction constraints.

1. The author constrains. At the first, I used the baseline classifier to predict labels. Then sum of the confidence values is assigned to the specific author. If the sum is positive,then the value is positive one. If the sum is negative, the value is negative one.
2. The user-interaction constraints. As a global model, stance classification is a sequence labeling task. Each of the training sequence corresponds to a post sequence. (Anand et al., 2011)

The classifier is ChainCRF model using the PyStruct, which is a Structured learning package in Python. The accuracy of global model will be discussed in the section 3.

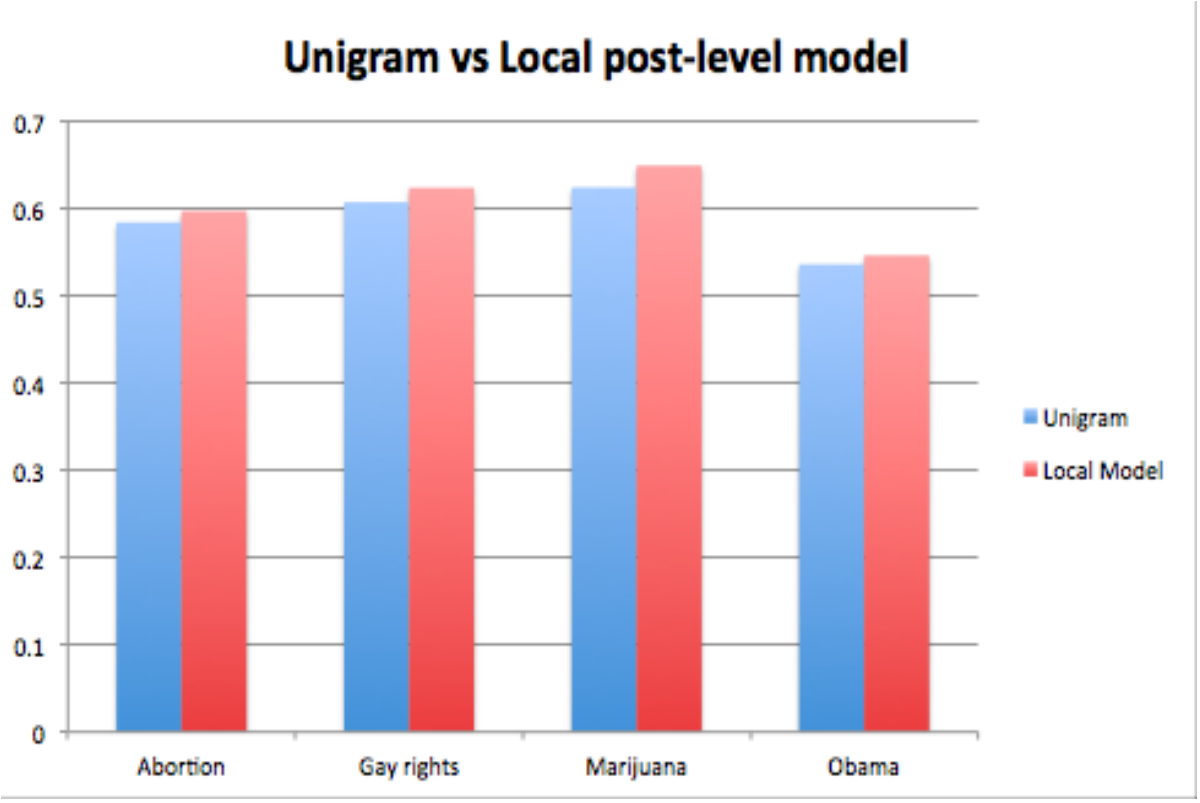
3 Accuracy

3.1 5-fold cross validation

In the project, I use 5-fold validation to test the accuracy. The fold sequence is the same as provided by the data sets. Every time, 4 of them are used for training and the another one is used for the testing. This method was used both in local model and global model.

3.2 Unigram vs Local model

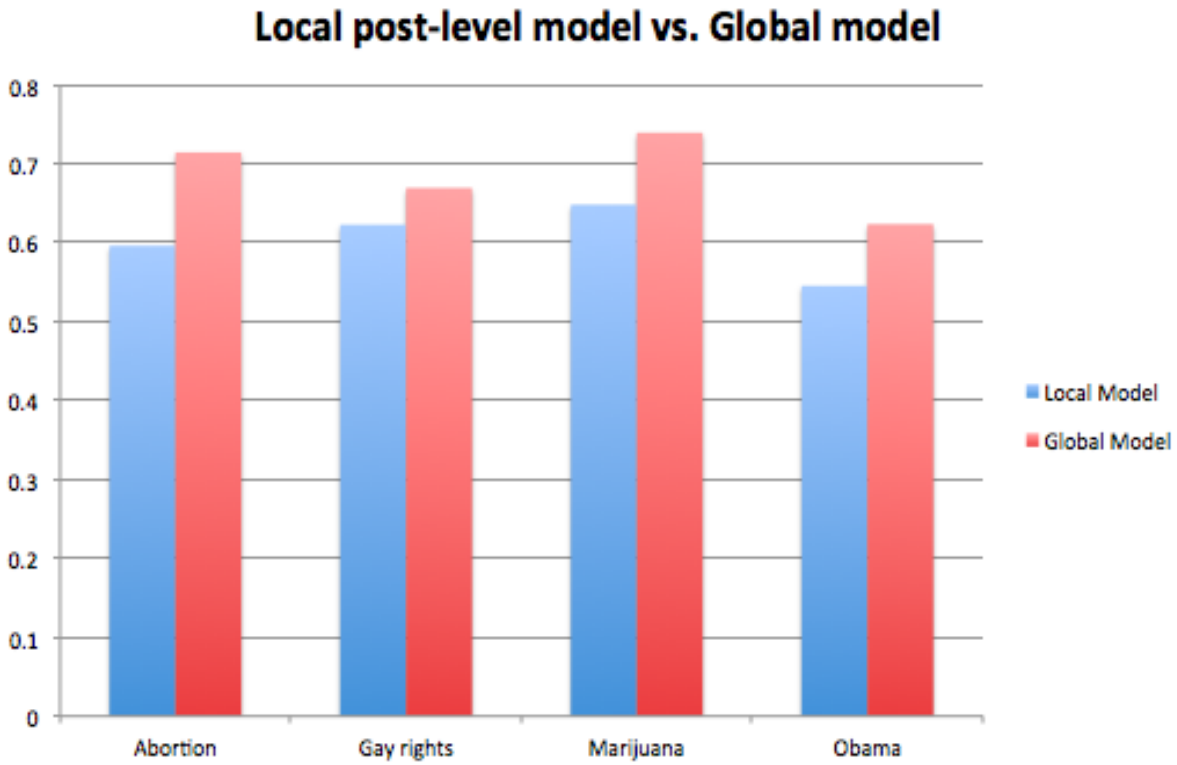
In this project, bag of the word model is the baseline. After add the features set from the local model, the accuracy of each class improve a little bit. The following picture represent the accuracy.



The average accuracy of four topics by using unigram is 57.67%, which is slightly better than the random pick. The local post-level model improve the unigram model about 2.6%. The average accuracy of the four topics: abortion, gay rights, Marijuana and obama are 59.63%, 62.26%, 64.82%, and 54.52%. It does not improved a lot. There are several reasons for that. One reason is the local model add some noises feature. For instance, the length of the post. Also, from the paper we know that the context factors works not well. In that case, it only improve a little bit. Overall, it is better than the bag of word model. In the following section, I will talked about the accuracy of the global model.

3.3 Local model vs Global model

The following picture represent the accuracy of the global model and local model.



The overall accuracy of global model is 68.65%. The average accuracy of the four topics: abortion, gay rights, Marijuana and obama are 71.43%, 66.92%, 73.92%, and 62.34%. Comparing with the local model, it improved about 8%. It is a little bit lower than the accuracy rates that represented in the paper. There are several reasons. I believe the main reason is the baseline model. In my project, the local post-level model is a little bit worse than the model of Hasan and Ng. (2013) In that case, the overall score of the global model score is lower than it in the paper.

Overall, the global model is better than the local model. However, the accuracy is a little bit lower than the model of Hasan and Ng. (2013). In the future work, I will improve it.