CS 583: DATA MINING AND TEXT MINING

Report of Project 2

Xuelong Wang & Shuanxi Zhu

April 29, 2017

1 ABSTRACT

In this report, we investigate the role of the social media twitter during the 2012 U.S. Presidential Election. The objective of this study was to identify the polarities of two candidates using historical Twitter data. A total of 14,000 tweets mentioned two candidates were pulled out from Twitter; human-labels were given in three categories for each tweets; various supervised learning methods were applied to find the best classifier in classifying the sentiment in such tweets. The analyses showed that a voting method of combing three supervised learning methods performed the best, and the social media data do contain some useful messages to infer the election results.

2 INTRODUCTION

Sentiment analysis, also known as opinion mining, is a process that aims to explore peoples' attitudes towards written languages via data mining techniques. In this report we described a framework for analyzing Twitter data during the 2012 Presidential Election. The goal is to train a good classifier that can correctly identify the polarity of a tweet. To simplify the problem, we only considered if a tweet contained positive, neutral, or negative sentiment. And we assumed that each tweet can convey a single category of opinion.

In the next few paragraphs, we first described the data and techniques used in training the classifier in Section 3; then we presented the evaluation metrics associated with our research in Section 4; finally we concluded the findings we had during the analyses and discussed the challenges of applying this classifier to general election data in Section 5.

3 TECHNIQUE

3.1 DATA

Data Collection: Since there were two candidates back in 2012, two domains of data were created for the candidates. For both Obama and Romney, 7000 tweets were streamed and coded by human being into three categories. Noted that, while all the steps of analyses are the same, the two datasets are trained and evaluated separately due to the difference in domain knowledge.

Data Clean: Before doing any real analyses, we obtained some descriptive statistics of the data and did some adjustment to remove the inconsistency among the data. Since only 3 categories were defined in the project description, all the tweets with labels out of range were discarded in the data cleaning process. This decreased the data size to 5624 for Obama and 5648 for Romney.

Data Pre-processing: Since the tweet data is in text format, it needs to be transformed into the structured data. We deleted stop-words. Features used here are the weighted combination of unigrams and n-grams of the tweets. The size of bags of words were chosen empirically, but it also makes sense in statistics due to the limited length of tweets. If we put two or more words in one bag, there will be not enough features thus not enough abstracted information in the analytic set. Therefore, we consider to combine unigram and n-grams together. After we obtained the features and counted the frequency of features in each tweet, tf-idf was applied on the vectorized data.

3.2 CLASSIFICATION METHOD

Several classification methods were tried here. And all of them are implemented by the scikit-learn toolkit in Python.

Stochastic Gradient Descent: This method are used for grid search and variable selection. It has different loss functions which corresponding to different methods(e.g. hing = SVM, log = logistic regression). The reason to use this method is its also has a tuning parameter corresponding to model complexity. Since the vocabulary of this tweet data is very large, so we want to use that parameter to find the significant words only in order to prevent overfitting problem.

Naive Bayes Classification: As the most well-known and well-studied method in supervised learning, naive bayes classifier has been widely applied in sentiment analyses. Because of its simplicity and intuitiveness, naive bayes classifier is always treated as baseline for classification task. However, due to the strong assumption of the distribution and the conditional independence constraint in features, this method might not performance well when the samples are not representative of the general population. Additionally, compared with other popular methods such as logistic regression or SVM, naive classifier is not very efficient.

Multivariate Logistic Regression: Logistic regression, initially introduced to solve the prediction problems in categorical data, is inherently suitable to solve this problem. Since we have 3 categories to predict here, multivariate logistic regression was chosen here. Due to the large volume of features compared with lines of tweets, we used the L1 penalized method, which will force the feature estimations which are close to zero converge to zero more quickly. This method is computational efficient and designed to solve the high dimensional modeling problem.

Support Vector Machine: In general, Support Vector Machine(SVM) is a linear learning system that builds two-class. Due to its high precision in text mining task, we also applied this classifier on our data. Some different strategies were applied in order to adjust this method in 3-category case. It is called one-vs-the-rest(OVR) multilabel strategy. Instead of fitting one classifier, OVR fits one classifier for each class. In each classifier, the class is fitted against all the other class in the data. There are two main advantages of this method, one is that this method runs very fast; and the other is that the results will be very easy to understand.

4 EVALUATION

To evaluate the classification performance of the classifiers described above, we used 10-cross-fold validation while training the classifier. Additionally, the test tweets were provided by the professor. Within the test set, there are also two sheets of human-labeled data, corresponding to Obama and Romney, respectively.

4.1 and 4.2 are the results of 10-fold-cross-validation for Obama and Romney, respectively. Here we only showed the results for positive and negative class as they are usually contained more information during the election.

From the table 4.1, we can see that SVM and logistic regression both perform better compared with naive bayes classifier.

Algorithms	Positive class			Negative class			Average
	precision	recall	F1-score	precision	recall	F1-score	
MultinominalNB	0.46	0.64	0.54	0.51	0.52	0.52	0.48
Logistic Regression	0.67	0.51	0.61	0.59	0.63	0.61	0.59
SVM	0.64	0.60	0.62	0.60	0.63	0.62	0.60

Table 4.1: 10-fold-cross-validation Results for Obama

For Romney's case, we can see that SVM performs the best based on the results in table 4.2.

Algorithms	Positive class			Negative class			Overall
	precision	recall	F1-score	precision	recall	F1-score	
Naive Bayes	0.28	0.48	0.35	0.62	0.43	0.51	0.42
Logistic Regression	0.60	0.26	0.36	0.58	0.86	0.70	0.57
SVM	0.56	0.41	0.47	0.64	0.77	0.70	0.58

Table 4.2: 10-fold-cross-validation Results for Romney

In general, Obama data gives better results compared with Romney. This makes sense as Obama data is more balanced and it is known that both logistic regression and SVM are sensitive to the balance of the data.

Improvement: After applied the penalty on complexity of each model used above and combination of the three classifiers, the results are improved by 2-5%. the details of the result of 10-fold cross validation is in the our directory. We think the penalty on the complexity should help our model predicts better on the test data.

The following table gives the results on the test dataset.

Candidate	Positive class			Negative class			Accu
	precision	recall	F1-score	precision	recall	F1-score	
Obama	0.62	0.53	0.57	0.57	0.65	0.60	0.58
Romney	0.66	0.49	0.6	0.68	0.79	0.73	0.64

Table 4.3: Classification Results on Test Data

5 CONCLUSION

In the report, we used a set of classifier techniques to predict the polarity of tweets. The results have showed that the best performance is achieved by combing three classifier together. This experiment also gives guidance of sampling strategy in data collection. In order to obtain a good classifier for sentiment prediction in tweet data, we need to make the data as balance as possible. On the other hand, as the golden standard in classification, human label should be as precise as possible to avoid any mistakes ore bias introduced into the classifier. Based on the data, we can also see that the social media does shed light on the results of Presidential Election.

REFERENCES

[1] LIU, BING. (2013). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.*, Berlin: Springer, 2013. Print.