Utrecht University

Data Mining: Assignment 2

TEXT CLASSIFICATION FOR THE DETECTION OF OPINION SPAM

Ana Borovac
6584446

Argyro (Iro) Sfoungari
6528015

October, 2018

Contents

Abstract

1 Introduction

Nowadays email filters are able to classify spam and not spam emails quite successfully. Imagine now that you have a web site with hotel reviews and your goal is to offer the most truthful reviews but you can not check every single review. Therefore you would like to have a mechanism which is going to help you to achieve your goal.

In this assignment we tried to solved the above problem with 4 methods; naive Bayes, logistic regression, classification tree and random forests. Before we started, we analysed the work that has been done already (section 2). Next, we did some data preprocessing, it is described in the section 3. Used methods are explained in the section 4 and analysis of the results (section 5) is in the section 6.

2 Related work

3 Data

Our data consists of 400 negative deceptive and 400 negative truthful hotel reviews that have been collected by Myble Ott and others ([1], [2]).

3.1 Data preprocessing

Before modeling we did some data preprocessing in order to get better models for predictiong if a review is fake or real. First step in data preprocessing was to remove punctuation marks, after that we made every letter lower case, we also removed stopwords, numbers and excess whitespace. Next, we created a test and a training set. Because we wanted to use cross validation, we divided our data into 5 folds (each of size 160 samples - 80 fake reviews and 80 truthful reviews). So, 4 of the folds represented a training set, the remaining 5th fold was a test set. At that moment every unique word from the training reviews was a feature. A number of features was large therefore we removed the words as features which occur less than 5 % of training documents. Next, we created a new training set which consists all the features from the previous one and bigrams. We again removed bigrams that occur in lass that 5 % of training documents. At the end we had two training sets ready to be used.

4 Methods

We used different classifiers to model our task:

- naive Bayes (generative linear classifier) – subsection 4.1,
- logistic regression (discriminative linear classifier) – subsection 4.2,

- classification tree (flexible classifier) – subsection 4.3,
- random forests (flexible classifier) – subsection 4.4.

4.1 Naive Bayes

For naive Bayes we used the code from the lectures ()

4.2 Logistic regression

For logistic regression model we used `cv.glmnet` from `glmnet` library

4.3 Classification tree

For growing the classification tree we used a function `rpart` from R library also called `rpart`. Library also contains a function for pruning the tree, `prune`. We used it with complexity parameter `cp` equals to 0,001.

4.4 Random forests

We grew a random forest of 200 trees with the function `randomForest` (R library `randomForest`). We set `mtry` parameter to 6 which means that on each step the algorithm selects 6 random observed features.

5 Results

6 Analysis

7 Conclusion

References

[1] Myle Ott, Claire Cardie, and Jeffrey T. Hancock. Negative deceptive opinion spam. In *in HLT-NAACL*, 2013.

[2] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.