# Report

## Assignment 2: Text classification for the Detection of the Opinion Spam

## Abstract

[write here the results]

## Introduction

The habit of reviewing products and services on online platforms ( Tripadvisor, Yelp, Rate Your Music, …) is increasing very quickly. In parallel, according to (Cone, 2011) users are relying more and more on others’ reviews when it comes to make a decision. Due to that, not all the reviews are truthful, but some of them are deliberately manufactured to sound authentic by malicious people who want to gain more customers against competitors. We call these fictitious reviews DECEPTIVE OPINION SPAM. Furthermore, we distinguish between them depending on the general *sentiment* a review expresses: positive reviews express satisfaction about the product or the service, while negative reviews express disappointment. We are aware that this distinction excludes all the possible intermediate evaluations, so it’s a strong limit for our analysis (following the normal categorization of web reviews, we are considering only 5-star and 1-star or 2-star reviews).

Previous works have already tried to build automatic classifiers which help to find out deceptive reviews and showed that they perform better than chance and human judge. In particular, Ott et al. (2011) focused on positive deceptive reviews, while Ott et al. (2013) focused on negative deceptive reviews. An important contribution of these researches is the creation of two datasets, respectively of gold-standard deceptive positive reviews and gold standard deceptive negative reviews. Indeed, the biggest obstacle to analysis before these two papers was the absence of sample data (reviews in our case) which can represent in an unbiased manner the deceptive reviews, since it’s complicated even to identify (at a certainty level of 100%) deceptive reviews, for obvious reasons. Thanks to that, further analysis is now possible exploiting those data.

We are trying now to improve the performance of the linear classifiers (naïve Bayes and Support Vector Machine) of the two previous works exploring different possibilities via more flexible classifiers and trying to build the best models via hyperparameters tuning. Due to simplicity reasons, we will focus only on negative reviews, leaving the positive ones to future work.

Our experiment is then divided into four parts, and each part takes into account a model. In details, we have:

1. Naïve Bayes
2. Regularized logistic regression
3. Classification tree
4. Random forests

For each model, we consider two sets generated by the reviews:

1. Only the unigrams
2. Unigrams and bigrams (set of all couples of two consecutive words)

In order to allow the reader to reproduce the experiments, all the code can be found in the Appendix parts.

## The data

As introduced before, we are exploiting the work of Ott et al. (2013) and using their data. Truthful reviews were collected from the following reviewing websites: Expedia, Hotels.com, Orbitz, Priceline, Tripadvisor and Yelp. Although they can’t be considered gold standard data, Mayzlin et al. (2012) and Ott et al. (2012) suggest that deception rate among those data is acceptably small. On the other side, deceptive opinion spam has been generated using Amazon’s Mechanical Turk service. The quality of these reviews has been proved by showing the actual difficulty of human testers in identifying them.

In details, the dataset is composed only of negative reviews (800). Data is equally divided into deceptive opinion spam (400) and truthful reviews (400). Each of these two sets is then divided into 5 folders. We use folders 1:4 of both deceptive and positive reviews (640 reviews in total) as training set, while we use folder 5 (160) as test set.

The reviews regard 20 popular hotels of Chicago, so that each hotel has 20 truthful reviews and 20 deceptive reviews. The truthful reviews were sampled according to a log normal distribution fit to the lengths of the deceptive reviews, since truthful reviews are on average longer than deceptive reviews. For further information about the collecting procedure and about Mechanical Tusk, refer to (Ott et al., 2013), the original paper.

## Setup of the experiments

All the experiments have been conducted in R language. In order to do that, it’s necessary to install an environment which allows to run R code. We have used Rstudio. Some additional packages have been installed to exploit their functions in the analysis. The following commands should then be typed first in the command line.

install.packages("tm")

install.packages("entropy")

install.packages("randomForest")

install.packages("randomForest")

install.packages("rpart")

install.packages("rpart.plot")

install.packages("glmnet")

Data can be downloaded from the following website:

<https://myleott.com/op-spam.html>

Then the following variables have been created in the Rstudio workspace. The four variables represent respectively the corpus of deceptive reviews in the training set, the corpus of the truthful reviews in the training set, the corpus of deceptive reviews in the test set, the corpus of truthful reviews in the test set. Be careful to substitute in the commands the real address of the folder which contains the data (usually, the Download folder).

training.corpus.dec <- c("C:/--address of the folder --op\_spam\_v1.4/negative\_polarity/deceptive\_from\_MTurk/fold1", "C:/--address of the folder -- op\_spam\_v1.4/negative\_polarity/deceptive\_from\_MTurk/fold2", "C:/--address of the folder -- op\_spam\_v1.4/negative\_polarity/deceptive\_from\_MTurk/fold3", "C:/--address of the folder -- op\_spam\_v1.4/negative\_polarity/deceptive\_from\_MTurk/fold4")

training.corpus.true<- c("C:/ --address of the folder --op\_spam\_v1.4/negative\_polarity/truthful\_from\_Web/fold2", "C:/--address of the folder -- op\_spam\_v1.4/negative\_polarity/truthful\_from\_Web/fold2", "C:/--address of the folder -- op\_spam\_v1.4/negative\_polarity/truthful\_from\_Web/fold3", "C:/--address of the folder -- op\_spam\_v1.4/negative\_polarity/truthful\_from\_Web/fold4")

testing.corpus.dec <- "C:/--address of the folder -- op\_spam\_v1.4/negative\_polarity/deceptive\_from\_MTurk/fold5"

testing.corpus.true <- “C:/ --address of the folder -- op\_spam\_v1.4/negative\_polarity/truthful\_from\_Web/fold5"

First, with the use of Natural language processing, we clean the data. We use the tm package, to

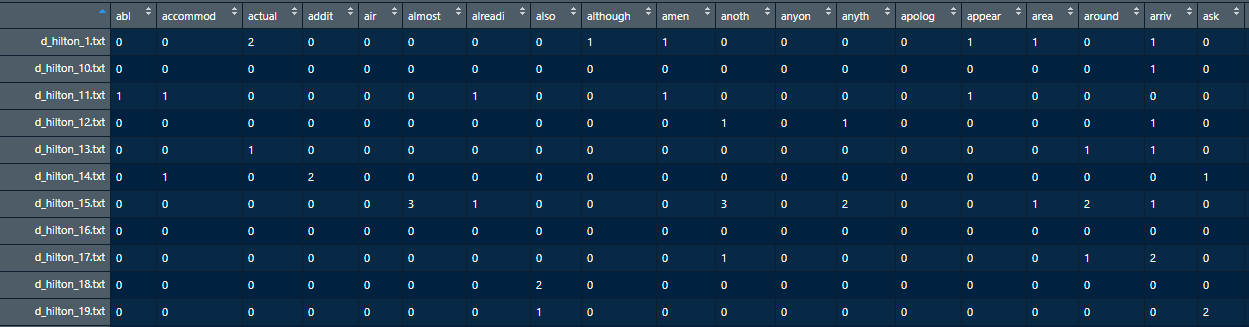
* Convert the entire text to lower case
* Remove numbers
* Remove whitespaces
* Remove punctuation
* Remove stopwords

In order to do that, we create a new Rscript with a function we call cleaning.function, to reuse it more times. The code of this function can be found in *Appendix1*. It unites the two matrices but does not mix deceptive and truthful reviews and returns a list of documents of words.

## Description of the experiments

### Naïve Bayes

For order purposes, we create a Rscript called Naïve Bayes and put all the code inside the function naïve.bayes.function. We clean the data using the cleaning function already mentioned, we extract all the unigrams, we remove all the words which appear in no more than 5% of the documents (“sparse terms”), we create a matrix. Here there is a part of it, just to show how data are represented in the dataset we are manipulating. We have the words as columns and the reviews as rows: each value (i,j) is 1 if the j-th word appears in the i-th review.



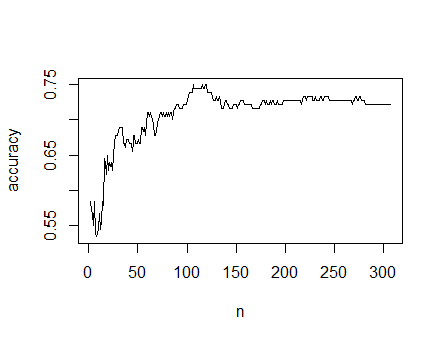
We do the same with bigrams. We have 307 unigrams and 12 bigrams. The bigrams are

"called front" "chicago hotel" "customer service" "even though"

"front desk" "got room" "hard rock" "hotel chicago"

"never stay" "room service" "stay hotel" "will never"

We train two models:

1. for the first, we compute the mutual information (MI) every feature (unigram) has and order the features according to the MI value. We then build 306 different models all with a different number of features, selecting each time the n best features according to the ordering, n ∈ [2,306]. We select the one with the highest accuracy, corresponding to n = 106. The following graph shows the accuracies depending on n.
2. For the second, we do the same as before including bigrams. We have then 318 different models, and n ∈ [2,318]. We have that the best n is 123. As an example, here there the best 10 features according to MI, and their values.

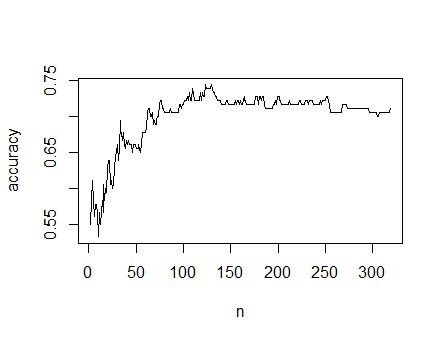
chicago location smell luxury hotel chicago

0.09752866 0.03632829 0.03275744 0.03246057 0.03018794

chicago hotel decided recently finally millennium

0.02797741 0.02588616 0.02519993 0.02244856 0.02215038

The following graph shows again the value of the accuracies depending on n.



Mutual information measures the reduction in uncertainty about X achieved by observing the value of Y (and vice versa). In our experiment we use it to correlate features and classes (spam – non spam). Filtering out some features according to mutual information can help in avoiding overfitting. Indeed, in both graphs the accuracy increases very fast for n which goes from 0 to 100 ( which means we are using too few features, we are underfitting), reaches a maximum between 100 and 150, and then slowly decreases ( which means we are overfitting).

The results are reported with the following contingency tables

1. Only unigrams

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Review is actually** | |
|  |  | Deceptive | Truthful |
| **Model predicts** | Deceptive | 66 | 11 |
| **the review** | Truthful | 14 | 69 |

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| 0,7500 | 0,8571 | 0,8250 | 0,8407 |

1. Both unigrams and digrams

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Review is actually** | |
|  |  | Deceptive | Truthful |
| **Model predicts** | Deceptive | 64 | 10 |
| **the review** | Truthful | 16 | 70 |

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| 0,7444 | 0,8649 | 0,8000 | 0,8312 |

Including bigrams does not improve the performances of the classifier (and it makes them worse indeed).

The code of this experiment can be found in *Appendix2*.

### Regularized Logistic Regression

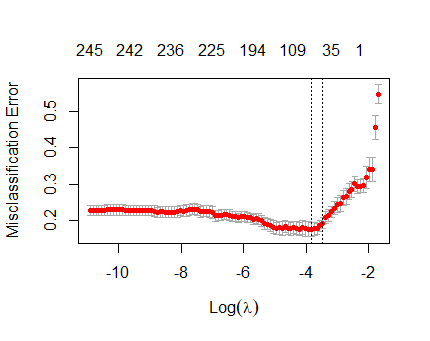
To prepare the data for this experiment we repeat the same procedure done in the previous experiment, for both unigrams and bigrams. We build two models, one with unigrams and the other with both unigrams and bigrams. We put all the code into a function, logistic.regression, that we call from the command line with the appropriate parameters.

For each model, the cv.glmnet package performs cross validation on the lambda hyperparameter of the regularization, but not on alpha. The default value of alpha is 1. No other cross validation to tune the hyperparameters is done.

The results are as follows.

1. Only with unigrams

The best lambda value is 0.02158487. This is the plot of different log lambda values against misclassification error.



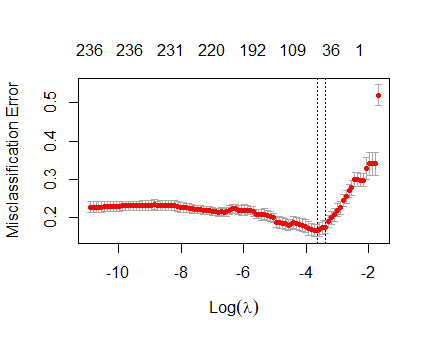
The following confusion matrix reports the performances.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Review is actually** | |
|  |  | Deceptive | Truthful |
| **Model predicts** | Deceptive | 58 | 8 |
| **The review** | Truthful | 22 | 72 |

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| 0,7222 | 0,8788 | 0,7250 | 0,7945 |

1. With both unigrams and bigrams

The best lambda value is 0.02599906. The following graphs are as previously.



|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Review is actually** | |
|  |  | Deceptive | Truthful |
| **Model predicts** | Deceptive | 58 | 8 |
| **The review** | Truthful | 22 | 72 |

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| 0,7222 | 0,8788 | 0,7250 | 0,7945 |

The lambda hyperparameter is a measure of how much we want to penalize high weights given to the features to predict the correct class. High weights indeed cause overfitting. The two plots show a common pattern: a high value of lambda produces a very simple model with almost all 0 weights which is unable to work correctly (underfitting). By decreasing lambda, we improve the performance until a maximum. If we do not stop increasing, we get a model with very high weight which is exposed to overfitting, causing the slow worsening of the performance.

The logistic regression classifiers have lower accuracies with respect to the naïve bayes classifiers, so we are induced to state that for this problem the generative linear model is a better choice than the discriminative linear model. However, the precision of the latter is a bit smaller than the precision of the former. A Mc Nemar test could help to find out if the differences are significant.

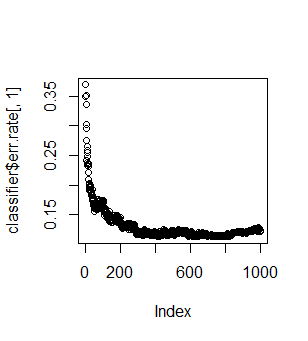
Including bigrams does not improve the performance (which is actually exactly the same).

The code of this experiment can be found in *Appendix3*.

### Classification Trees

### Random Forests

An out of bag evaluation has been carried out to find the best hyperparameters. The function RFtune has been used to find the best number of randomly selected features per each split. The number is 5 (while the standard number for the Random Forest packages is 17). Then, an out of bag evaluation has been carried out to find the best number of trees (between 1 and 1000). Hereunder, there is the graph which represents the results. “Index” stays for the number of trees used, while “classifier$err.rate[,1]” stays for the OOB error. The best number of trees is 415. To reproduce the experiment, see Appendix 5.



### Analysis 3 (classification trees)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Review is actually** | |
|  |  | Deceptive | Truthful |
| **Model predicts** | Deceptive | 42 | 16 |
| **The review** | Truthful | 38 | 64 |

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| 0,6625 | 0.7241 | 0,5250 | 0,6087 |

### Analysis 4 (random forests)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Review is actually** | |
|  |  | Deceptive | Truthful |
| **Model predicts** | Deceptive | 69 | 24 |
| **The review** | Truthful | 11 | 56 |

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| 0,7812 | 0,7420 | 0,8625 | 0,7977 |

## Conclusions

## Appendix 1: Cleaning function

#function used to clean the data

library(tm)

cleaning.function <- function(corpus.dec, corpus.true){

reviews.dec <-VCorpus(DirSource(corpus.dec,encoding="UTF-8"))

reviews.true<-VCorpus(DirSource(corpus.true,encoding="UTF-8"))

review.all<-c(reviews.dec,reviews.true)

#clean the data

review.all <- tm\_map(review.all, content\_transformer(tolower))

review.all <- tm\_map(review.all, removeNumbers)

review.all <- tm\_map(review.all, removePunctuation)

review.all <- tm\_map(review.all, stripWhitespace)

review.all <- tm\_map(review.all,removeWords,stopwords("english"))

return (review.all)

}

## Appendix 2: Code for Naïve Bayes analysis

#--------------------------libraries---------------------#

library(entropy)

naive.bayes.function <- function (training.corpus.dec, training.corpus.true, testing.corpus.dec, testing.corpus.true){

#training set

training.dtm <- cleaning.function(training.corpus.dec,training.corpus.true)

#extraction of unigrams

training.dtm.unigrams <- DocumentTermMatrix(training.dtm)

training.dtm.unigrams <- removeSparseTerms(training.dtm.unigrams,0.95)

training.dtm.unigrams <- as.matrix(training.dtm.unigrams)

#extraction of bigrams

BigramTokenizer <-function(x) unlist(lapply(ngrams(words(x), 2), paste, collapse = " "), use.names = FALSE)

training.dtm.bigrams <- DocumentTermMatrix(training.dtm,control = list(tokenize = BigramTokenizer))

training.dtm.bigrams <- removeSparseTerms(training.dtm.bigrams,0.95)

training.dtm.bigrams <- as.matrix(training.dtm.bigrams)

#training set with both unigrams and bigrams

training.labels <- c(rep(0,320),rep(1,320))

training.dtm <- cbind(training.dtm.unigrams, training.dtm.bigrams)

###################################################################################

#test set

test.dtm <- cleaning.function (testing.corpus.dec,testing.corpus.true)

#unigrams

test.dtm.unigrams<- DocumentTermMatrix(test.dtm,list(dictionary=dimnames(training.dtm.unigrams)[[2]]))

test.dtm.unigrams <- as.matrix(test.dtm.unigrams)

#bigrams

test.dtm.bigrams<- DocumentTermMatrix(test.dtm,list(dictionary=dimnames(training.dtm.bigrams)[[2]]))

test.dtm.bigrams <- as.matrix(test.dtm.bigrams)

#test set with both unigrams and bigrams

test.labels <- c(rep(0,80),rep(1,80))

test.dtm <- cbind(test.dtm.unigrams, test.dtm.bigrams)

########################################################################################

#feature selection (with mutual information, only for unigrams)

training.dtm.unigrams.mi <- apply(training.dtm.unigrams,2,function(x,y){

mi.plugin(table(x,y)/length(y))},training.labels)

training.dtm.unigrams.mi.order <- order(training.dtm.unigrams.mi,decreasing = T)

#feature selection ( with mutual information)

training.dtm.mi <- apply(training.dtm,2,function(x,y){

mi.plugin(table(x,y)/length(y))},training.labels)

training.dtm.mi.order <- order(training.dtm.mi,decreasing = T)

###########################################################################################

#print

print(dim(training.dtm.bigrams))

print(dim(training.dtm.unigrams))

print(colnames(training.dtm.bigrams))

print(dim(training.dtm)) ## just to check

print(training.dtm.mi[training.dtm.mi.order[1:10]])

#first model ( with feature selection according to mutual information)(only unigrams)

accuracies.unigrams.mi.models <- sapply(c(2:307), function (num.features){

model.mi <-train.mnb(training.dtm.unigrams[,training.dtm.unigrams.mi.order[1:num.features] ], training.labels)

predictions.mi <- predict.mnb(model.mi , test.dtm.unigrams[,training.dtm.unigrams.mi.order[1:num.features]])

conf.mat <- table (predictions.mi ,test.labels)

return (sum(diag(conf.mat))/180)

} )

accuracies.unigrams.mat <- cbind (accuracies.unigrams.mi.models, c(2:307))

accuracies.unigrams.best.n <- accuracies.unigrams.mat[which.max(accuracies.unigrams.mat[,1]), 2]

print(accuracies.unigrams.mat)

print(accuracies.unigrams.best.n)

plot(accuracies.unigrams.mat[,2] ,accuracies.unigrams.mat[,1], xlab = "n", ylab = "accuracy", type = "l")

model.unigrams.mi <-train.mnb(training.dtm.unigrams[,training.dtm.unigrams.mi.order[1:accuracies.unigrams.best.n] ], training.labels)

predictions.unigrams.mi <- predict.mnb(model.unigrams.mi , test.dtm.unigrams[,training.dtm.unigrams.mi.order[1:accuracies.unigrams.best.n]])

print(table (predictions.unigrams.mi ,test.labels))

#second model ( with feature selection according to mutual information)(both unigrams and bigrams)

accuracies.mi.models <- sapply(c(2:319), function (num.features){

model.mi <-train.mnb(training.dtm[,training.dtm.mi.order[1:num.features] ], training.labels)

predictions.mi <- predict.mnb(model.mi , test.dtm[,training.dtm.mi.order[1:num.features]])

conf.mat <- table (predictions.mi ,test.labels)

return (sum(diag(conf.mat))/180)

} )

accuracies.mat <- cbind (accuracies.mi.models, c(2:319))

accuracies.best <- accuracies.mat[which.max(accuracies.mat[,1]), 2]

print(accuracies.mat)

print(accuracies.best)

plot(accuracies.mat[,2] ,accuracies.mat[,1], xlab = "n", ylab = "accuracy", type = "l")

model.mi <-train.mnb(training.dtm[,training.dtm.mi.order[1:accuracies.best] ], training.labels)

predictions.mi <- predict.mnb(model.mi , test.dtm[,training.dtm.mi.order[1:accuracies.best]])

print(table (predictions.mi ,test.labels))

}

#Training function for Naive Bayes

#labels = classes

train.mnb <- function (dtm,labels) {

call <- match.call()

V <- ncol(dtm) #vocabulary

N <- nrow(dtm) #number of documents

prior <- table(labels)/N

labelnames <- names(prior)

nclass <- length(prior)

cond.probs <- matrix(nrow=V,ncol=nclass)

dimnames(cond.probs)[[1]] <- dimnames(dtm)[[2]]

dimnames(cond.probs)[[2]] <- labelnames

index <- list(length=nclass)

for(j in 1:nclass){

index[[j]] <- c(1:N)[labels == labelnames[j]]

}

for(i in 1:V){

for(j in 1:nclass){

cond.probs[i,j] <- (sum(dtm[index[[j]],i])+1)/(sum(dtm[index[[j]],])+V)

#Laplace smoothing

}

}

x <- list(call=call,prior=prior,cond.probs=cond.probs)

return (x)

}

predict.mnb <- function (model,dtm) {

classlabels <- dimnames(model$cond.probs)[[2]]

logprobs <- dtm %\*% log(as.matrix(model$cond.probs))

N <- nrow(dtm) #number of documents to classify

nclass <- ncol(model$cond.probs) #number of classes

logprobs <- logprobs+matrix(nrow=N,ncol=nclass,log(model$prior),byrow=T)

x <- classlabels[max.col(logprobs)]

return (x)

}

## Appendix 3: Code for regularized logistic regression analysis

#--------------------------libraries---------------------#

library("glmnet")

logistic.function <- function (training.corpus.dec, training.corpus.true, testing.corpus.dec, testing.corpus.true){

#training set

training.dtm <- cleaning.function(training.corpus.dec,training.corpus.true)

#extraction of unigrams

training.dtm.unigrams <- DocumentTermMatrix(training.dtm)

training.dtm.unigrams <- removeSparseTerms(training.dtm.unigrams,0.95)

training.dtm.unigrams <- as.matrix(training.dtm.unigrams)

#extraction of bigrams

BigramTokenizer <-function(x) unlist(lapply(ngrams(words(x), 2), paste, collapse = " "), use.names = FALSE)

training.dtm.bigrams <- DocumentTermMatrix(training.dtm,control = list(tokenize = BigramTokenizer))

training.dtm.bigrams <- removeSparseTerms(training.dtm.bigrams,0.95)

training.dtm.bigrams <- as.matrix(training.dtm.bigrams)

#training set with both unigrams and bigrams

training.labels <- c(rep(0,320),rep(1,320))

training.dtm <- cbind(training.dtm.unigrams, training.dtm.bigrams)

#test set

test.dtm <- cleaning.function (testing.corpus.dec,testing.corpus.true)

#unigrams

test.dtm.unigrams<- DocumentTermMatrix(test.dtm,list(dictionary=dimnames(training.dtm.unigrams)[[2]]))

test.dtm.unigrams <- as.matrix(test.dtm.unigrams)

#bigrams

test.dtm.bigrams<- DocumentTermMatrix(test.dtm,list(dictionary=dimnames(training.dtm.bigrams)[[2]]))

test.dtm.bigrams <- as.matrix(test.dtm.bigrams)

#test set with both unigrams and bigrams

test.labels <- c(rep(0,80),rep(1,80))

test.dtm <- cbind(test.dtm.unigrams, test.dtm.bigrams)

#first model (only unigrams)

reviews.glmnet.unigrams <- cv.glmnet(training.dtm.unigrams,training.labels, family="binomial",type.measure="class")

print(coef(reviews.glmnet.unigrams,s="lambda.min"))

print(reviews.glmnet.unigrams$lambda.min)

plot (reviews.glmnet.unigrams)

reviews.logreg.pred.unigrams <- predict(reviews.glmnet.unigrams,

newx=test.dtm.unigrams,s="lambda.min",type="class")

print(table(reviews.logreg.pred.unigrams,test.labels))

#second model (with bigrams)

reviews.glmnet <- cv.glmnet(training.dtm,training.labels, family="binomial",type.measure="class")

print(coef(reviews.glmnet,s="lambda.min"))

print(reviews.glmnet$lambda.min)

plot(reviews.glmnet)

reviews.logreg.pred <- predict(reviews.glmnet,

newx=test.dtm,s="lambda.min",type="class")

print(table(reviews.logreg.pred.unigrams,test.labels))

}

## Appendix 4: Code for classification tree

library("glmnet")

logistic.function <- function (training.corpus.dec, training.corpus.true, testing.corpus.dec, testing.corpus.true){

#training set

training.dtm <- cleaning.function(training.corpus.dec,training.corpus.true)

training.dtm <- DocumentTermMatrix(training.dtm)

training.dtm <- removeSparseTerms(training.dtm,0.95)

training.dtm = as.matrix(training.dtm)

training.labels <- c(rep(0,320),rep(1,320))

#test set

test.dtm <- DocumentTermMatrix(cleaning.function (testing.corpus.dec,testing.corpus.true),list(dictionary=dimnames(training.dtm)[[2]]))

test.dtm = as.matrix(test.dtm)

test.labels <- c(rep(0,80),rep(1,80))

reviews.glmnet <- cv.glmnet(training.dtm,training.labels,

family="binomial",type.measure="class")

print (coef(reviews.glmnet,s="lambda.1se"))

reviews.logreg.pred <- predict(reviews.glmnet,

newx=test.dtm,s="lambda.1se",type="class")

table(reviews.logreg.pred,test.labels)

}

## References

[1] Myle Ott, Yejin Choi, Claire Cardie and Jerey T. Hancock, Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th meeting of the association for computational linguistics, pp. 309-319,2011.

[2] Myle Ott, Claire Cardie and Jerey T. Hancock, Negative deceptive opinion spam. Proceedings of NAACL-HLT 2013, pp. 497-501, 2013.

[3] Cone. 2011. 2011 Online Influence Trend Tracker.