Classification of review sentiments

Rediet Tilahun Desta
Department of Electrical Engineering
rtilahun@princeton.edu

Abstract

In this age where we mostly buy products be they clothes or music, we would like to get the opinions of others before making the decision. However going through all reviews can be time consuming. Companies that faciliate this trading process would like to present their customers with a quick and easy way to gauge the overall response to the product; the first step of which will be analyzing the sentiments of each reviews. In this assignment, I attempt to find a good sentiment classifier which would classify reviews as either positive or negative, using a data set of 3000 reviews. I employed Natural Language Processing(NLP) techniques to represent the data as a bag-of-words which were then used to train six types of classifiers. Multinomial and bernoulli naive bayes classifiers, decision trees and random forest classifiers have high precision and recall, whereas support vector machines have a significant advantage in precision while performing significantly poorly in recall. K-nearest neighbours classifier has relatively good precision but performs poorly on recall.

1 Introduction

Sentiment analysis used to be a very human task in the past. However in recent years, with the growth of online companies, the need for automated forms of sentiment analysis were found to be of most importance. Online companies like Amazon, Facebook, Twitter and Yelp are increasingly becoming dependent on machine learning algorithms to classify the sentiment of the billions of textual inputs they receive daily.(The impact of online user reviews on hotel room sales,Ye et al.) In this assignment, I am interested in learning which type of classifiers perform well in classifying the sentiments of online reviews. Moreover, I want to gain the understanding of why certain classifiers perform well while others less so.

I evaluate 6 types of classifiers in this report. In the feature extraction process, I chose unigrams and then in the feature selection component selected those unigrams that appear more than 5 times in the 2400 data set. Other feature selection techniques were also used and then compared as to their impact on classifier performance such as computational speed, precision and recall.

2 Related Work

The first indepth analysis of the field of sentiment analyis and classification was done by Pang et. al in 2008.[**?**] In trying to understand the sentiment expressed in a document one can dissect the problem to understanding the sentimental value of the words in that document and the sentences that make up that document. Moreover, one needs to distinguish between words and sentences that matter and those that don't matter. In the literature, these distinctions are known as subjectivity classification.[**?**] Tang clearly states that, "Document sentiment classification and opinion extraction have often involved word sentiment classification techniques."[**?**](A survey on sentiment detection of reviews, Tang et al.)For example, Pak et. al. use a sentiment classifier that is "based on the multinomial Naive Bayes classifier that uses N-gramand POS-tags [part of speech tagging] as

features."[**?**](Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Pak et. al.)Zhang et al. found that, "accuracy is influenced by interaction between the classification models and the feature options...Character-basedbigrams are proved better features than unigrams and trigrams in capturing Cantonese sentimentorientation."[**?**](Sentiment classification of Internet restaurant reviews written in Cantonese, zhang et. al.)

2.1 Data processing

I initially downloaded two data sets marked train.txt and test.txt. train.txt contained 2400 reviews, each marked with a 1 or a 0 by a human agent: 1 indicating a positive review and 0 a negative review. test.txt contained 600 distinct reviews from the same source. By using python's NLTK library, we were able to tokenize each document, convert it lower case words then remove words that were stop words, lemmatize and then stem it. After that we also took out words that did not appear more than 5 times in the whole data set. This resulted in a vocabulary set of size 541 words. These 541 words were used as features of each document in a bag-of-words format; their presence in the review being marked by 1.

2.2 Classification Methods

I used 6 different classification methods using Python's SciKitLearn libraries.[**?**]

1. *Multinomial Naive Bayes classifier* (MNB): using multinomial implementation
2. *Bernoulli Naive Bayes classifier* (BNB): using bernoulli implementation
3. *Support vector machine* (SVM):
4. *Decision tree* (DT):
5. *Random forest* (RF): using 100 trees
6. *K-nearest neighbors* (KNN): using ten nearest neighbors and the "KDTree" algorithm

Each classifier was first trained using the dataset found in train.txt. Python's SciKitLearn library provides functions that fit the features of dataset with the predetermined classes for each classifier type. However to use random forest classifiers and k-nearest neighbours classifiers, we needed to fit their hyperparameters. Hence, we used 10-fold cross validation on the training dataset and determeined the n-value which gave us the best accuracy. These n-values were then used as hyperparameters for classifiying the test data set using a random forest classifier or k-nearest neighbours.

2.3 Evaluation

We evaluated the six classification methods we selected using five different metrics. We first combined the test and training datasets.Then we used 10-fold cross validation technique on the dataset to iteratively train and test portions of overall data set. This technique guarantees that each document in the data set will be used as a testing document once. After the iteration completes we began to compare the predicted class with the predetermined class of the whole combined data set using the five aforementioned metrics. These metrics are accuracy, precision, recall, F_1 score, and time.

Below is brief definition of the five metrics:

1. $Accuracy = \frac{Numberofaccuratepredictions}{Numberofdocuments}$
2. $Precision = \frac{TruePositive}{TruePositive + FalsePositive} = 1 - FalseDiscoveryRate$
3. $Recall = \frac{TruePositive}{TruePositive + FalseNegative}$
4. $F_1\ score = 2 * \frac{precision \times recall}{precision + recall} = 2 * \frac{TruePostive}{2 * TruePositive + FalsePositive + FasleNegative}$
5. *Time* : $computationalspeedoftheprogramtofitthedatainseconds.$

3 Spotlight Classifier: Naive Bayes Classifier

Naive Bayes classifier is a very popular classifier used in text classification tasks especially as a baseline tool. It is fairly simple conceptually compared to other more sophisticated classifiers however when it comes to performance in text classification it is quite comparable even outperforming some.[**?**](Tackling the Poor Assumptions of Naive Bayes Text Classifiers, Rennie)

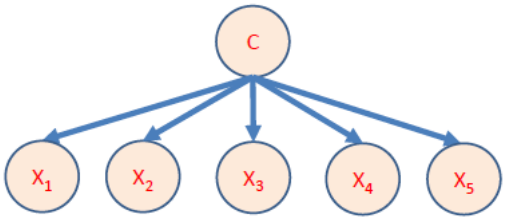


Figure 1: The caption

The basis of a naive bayes classifier is the conditional probability, which asks a very simple question: what is the probability that a certain random variable in our case a document, is either a positive or a negative review, given that it contains a certain sequence of words. Given a feature set of $F = \{\boldsymbol{x_1}, \boldsymbol{x_2}...\boldsymbol{x_N}\}$, and a class set of $C = \{\boldsymbol{c_1}, \boldsymbol{c_2}...\boldsymbol{c_N}\}$, we want to find

$$\arg\max_{c} P(C = c|F)$$

Using Bayes theorem, we know that,

$$P(C|x) = \frac{P(C)P(x|C)}{P(x)}$$

In our case the feature set contains more than one element. Hence we need to extend Bayes theorem so that C is conditioned on all the elements of the feature set and their values. Using the chain rule of conditional probability we then can further expand the equation. However, that expanded equation can become very tractable by naively assuming that all the elements in the feature set are independent of each other. Hence the previous equation can be approximated to,

$$\alpha\ P(C|x) = \frac{P(C)\prod_{i=a}^{b} P(x_i|C)}{P(x)}$$

However, that expanded equation can become very tractable by naively assuming that all the elements in the feature set are independent of each other. Hence the previous equation can be approximated to, Using the chain rule of conditional probability we then can further expand the equation.

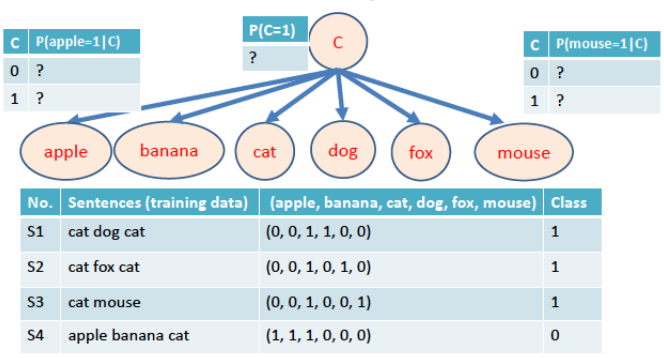
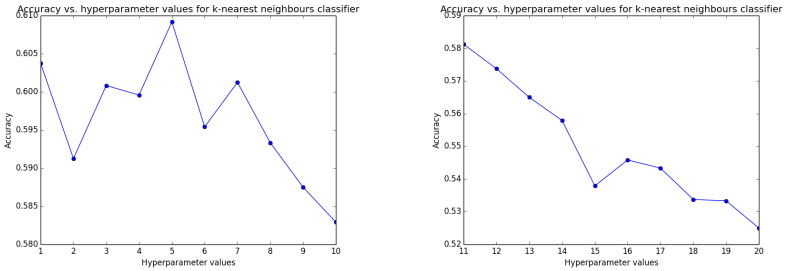


Figure 2: The caption

4 Results

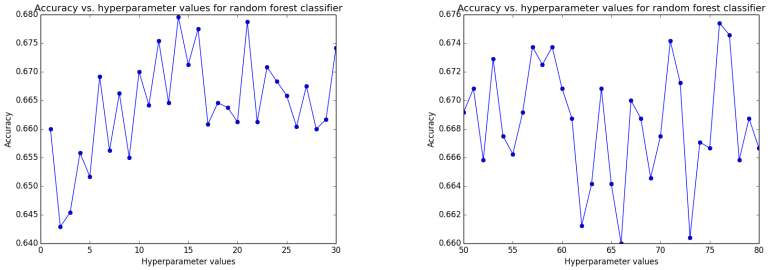
4.1 Evaluation results

Classifier	Accu	Prec	Recall	F_1	Time (s)
MNB	0.693	0.690	0.699	0.694	0.431
BNB	0.694	0.696	0.689	0.693	0.557
SVM	0.686	0.676	0.713	0.694	14.431
DT	0.677	0.671	0.696	0.6832	14.409
RT	0.679	0.667	0.713	0.689	14.425
KNN	0.650	0.699	0.526	0.600	21.320



(a) A subfigure (b) A subfigure

Figure 3: A figure with two subfigures



(a) A subfigure (b) A subfigure

Figure 4: A figure with two subfigures

4.2 Computational speed

However, that expanded equation can become very tractable by naively assuming that all the elements in the feature set are independent of each other. Hence the previous equation can be approximated to, Using the chain rule of conditional probability we then can further expand the equation.

As can be seen from support vector machines classifier and k-nearest neighbours classifiers take a significant amount of time to train. The time given is the amount of time taken to train a 3000 document data set using 10-fold cross validation technique. The fastest classifiers are as expected Multinomial Naive Bayes and Bernoulli Naive Bayes. They are the fastest amongst the classifiers because the number of parameters required is linear to the number of features in a learning problem. "Moreover, Maximum-likelihood training can be done by evaluating a closed-form expression,[1]which takes linear time."(Wikipedia)[**?**]

5 Discussion and Conclusion

Acknowledgments