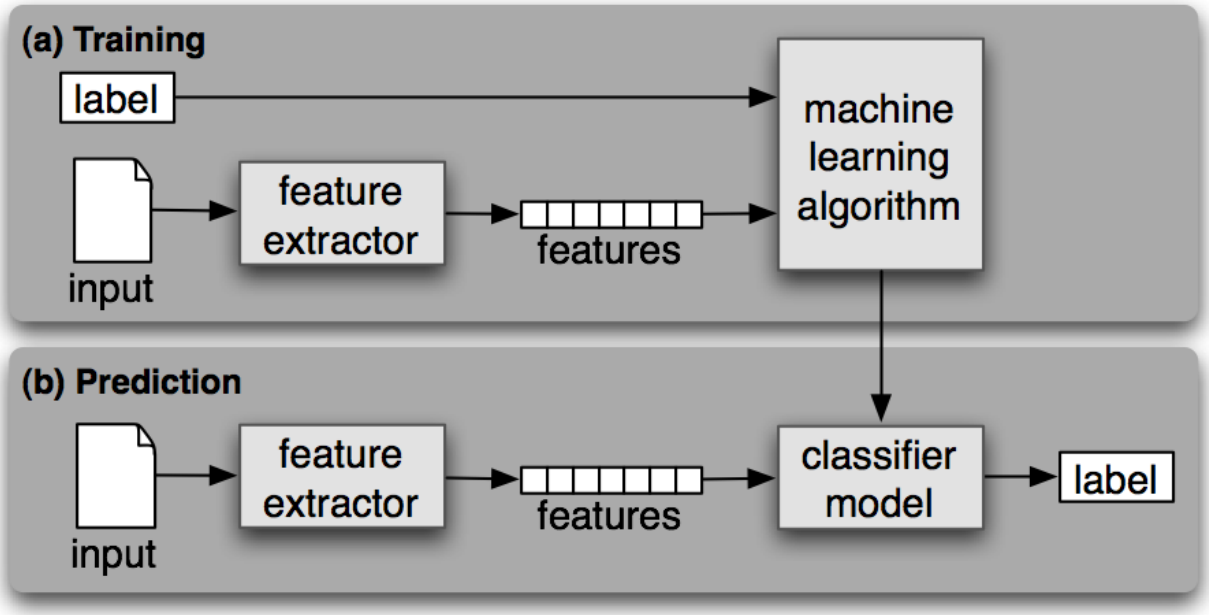
Abstract

The sample data set given has files with movie review comments which are already tagged as positive or negative based on the review comments. Create a machine learning model or classifier using naïve bayes technique that can be used to classify new movie reviews.

Requirements

1. Feature extraction
2. Identify the training, validation and test data
3. Create and train the model using training data
4. Validate the model using validation data
5. Verify the model with test data
6. Calculate the accuracy of classification

Design



4. code

```
#!/usr/bin/python


import os, sys

import nltk

import random

file_paths1=[]

file_paths=[]

count = {}


DIR = r"C:\Python Projects\pos"

for root,directories,files in os.walk(DIR):

  for filename in files:

    filepath=os.path.join(root,filename)

    file_paths.append(filepath)


all_words=[]

lnames=[]

lpos=[[[],'pos']]

for p in file_paths:

  lnames=open(p,'r').read().split()

  lpos.append([lnames,'pos'])

  for w in lnames:

    all_words.append(w)
```

```
DIR1 = r"C:\Python Projects\neg"
for root,directories,files in os.walk(DIR1):
  for filename in files:
    filepath1=os.path.join(root,filename)
    file_paths1.append(filepath1)

for q in file_paths1:
  lnames=open(q,'r').read().split()
  lpos.append([lnames,'neg'])
  for w in lnames:
    all_words.append(w)

random.shuffle(lpos)
print(len(all_words))

word_features=list(all_words)[:2000]

def document_features(document):
   document_words = set(document)
   features = {}
   for word in word_features:
      features['contains(%s)' % word] = (word in document_words)
   return features
```

```
featuresets=[(document_features(d),c) for (d,c) in lpos]

train_set,test_set=featuresets[5:],featuresets[:5]


classifier = nltk.NaiveBayesClassifier.train(train_set)

print("Accuracy in %:")

print (nltk.classify.accuracy(classifier, test_set)*100)

classifier.show_most_informative_features(5)
```

Conclusion & Challenges

The entertainment industry requires new and better ways to target specific users with certain features. This project is exploring the possibility of classifying the movie review corpus as positive or negative to help enable make better decisions for target users, using data mining techniques.

References

- (Alpaydin, 2004): For a general introduction to machine learning.
- (Manning & Schutze, 1999): techniques for language problems.
- (Manning, Raghavan, & Schutze, 2008):Naïve Bayes for classifying text.