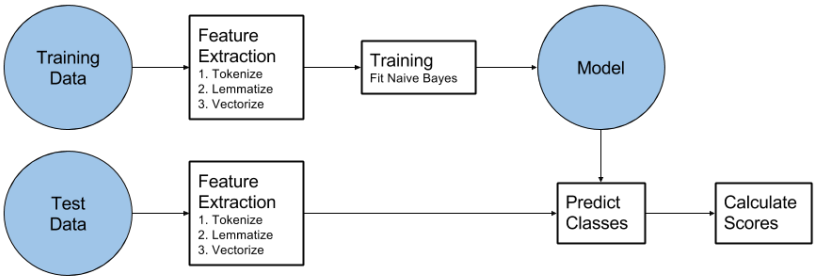
DM-2583 - Big Data in Media Technology
Lab 1 - Sentiment Analysis using a Naive Bayes Classifier

Charles Hamesse, Carl Holmqvist, Isak Stensö, Philip Stiff

Abstract

In this project, we use a dataset containing book reviews to train a natural language classifier for sentiment analysis in further test it on hotel reviews.

1 Process diagram



2 Feature extraction

The short texts in the data are tokenized, lemmatized and finally vectorized. By tokenizing the texts we can analyze on individual words. By lemmatizing the words, we get the meaning of the word, and have it in a form so that it can be compared to the same word but inflected differently. By vectorizing them, we turn the text strings into vectors, representing how many time each lemmatized word appears in every string.

Here's an example of the feature extraction process for the sentence "Brokeback Mountain was really depressing":

1. Tokenize: the sentence becomes a bag of words: Brokeback, Mountain, was, really, depressing
2. Lemmatize: Brokeback, Mountain, be, really, depress
3. Vectorize: Term-document matrix representing the frequencies of every dictionary word in the sentence. Brokeback, Mountain, be, really, depress are set to 1, all others are set to 0.

3 Model implementation

We used scikit-learn's MultinomialNB model.

4 Results

We got 135/200 correct predictions on the test data set using our trained model, representing 67.5% accuracy. Both the precision and the recall values for the whole data set were roughly 68%.

5 Evaluation

The results are affected by the subjectivity in annotating the test data ourselves. Since different people might annotate the data differently the results will be affected by the annotator.

68% is better than randomly guessing, which should (with an unknown test data set) yield 50% accuracy. However, the results are not overwhelmingly good. One reason could be that the training data is small. Also, the training data contains movie reviews, while the test data contains hotel reviews. It is possible that the different subject matters skew the results. Also, we are analyzing every word independently, regardless of its context. Using bi-grams could potentially yield better results.