NATURAL LANGUAGE PROCESSING

I5GIC

Mini Project 2 – Text Classification

Given a corpus of product reviews, build a text classifier to predict if an input review is positive or negative. The corpus includes 2 text files:

- positive-reviews.txt contains 22,936 lines of reviews (one review per line) considered to be positive
- negative-reviews.txt contains 22,932 lines of negative reviews

You can use the top 80% lines of each file as your training set. The rest will be used as a test set. You can implement any classification model of your choice. It would be great to implement multiple models and compare their results. Accuracy rate should be used as an evaluation metric.

The following features should be extracted from each review:

- count of positive words
- count of negative words
- 1 if the review contains the word 'no' or 0 otherwise
- count of 1st and 2nd pronouns ('I', 'me', 'my', 'you', 'your')
- 1 if '!' can be found in the review or 0 otherwise
- log(length of the review)

You can add more features that you find helpful in improving the classifier performance. You can find two additional text files: positive-words.txt and negative-words.txt which contain 2,006 positive words and 4,783 negative words respectively[1].

Write a report (max 2 pages) about the features used, the architecture of the models implemented as well as the results obtained from the experiments.

[1] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge; Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle Washington, USA