


CA640

PRACTICUM APPROVAL

Comparing Machine learning and
Lexicon-based approach for
sentiment analysis



Student Details

Names: Rohit Nair & Ashwin Bharadwaj

Student IDs: 20210378 & 20210149

Emails: rohit.nair2@mail.dcu.ie & ashwin.sridhar3@mail.dcu.ie

Major: Data Analytics

Supervisor: Dr. Jennifer Foster

Date of Submission: 06/12/2020

Datasets

Three datasets from three different domains will be used:

- IMDB Movie Reviews
- Amazon product reviews
- Tweets

1. IMDB Movie Reviews Dataset:

Dataset is obtained from **Stanford Artificial Intelligence Laboratory**. There are 12,000 movie reviews each for testing and training the model, labelled as positive and negative.

URL: http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

Datasets

2. Amazon Reviews:

Amazon dataset has approximately 120,000 records and the ratings are on a scale of 5. These ratings will be converted to positive or negative to train the model. The dataset is obtained from **Stanford Network Analysis Platform**.

URL: <http://snap.stanford.edu/data/web-Amazon-links.html>

3. Twitter Data set:

The twitter dataset consists of 1.5 million tweets labelled with 1 and 0. 1 being positive and 0 being negative. This dataset was obtained from **Thinknook.com**.

URL: <http://thinknook.com/wp-content/uploads/2012/09/Sentiment-Analysis-Dataset.zip>

Experiment and Results

- Implement Lexicon-based approach and Naive Bayes algorithm on the datasets.
- The trained model will be able to identify the sentiment of users from the data provided. Performance of the 2 models will be compared for each dataset.
- The results will be evaluated by calculating the accuracy percentage of the model based on the difference in predicted sentiment and actual sentiment.

Comparison:

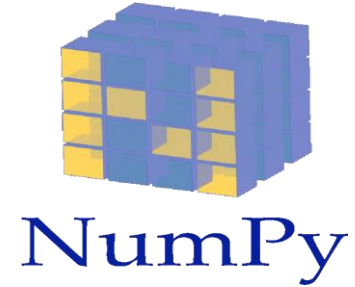
Naive - Bayes Algorithm:

- Is a probabilistic classifier based on applying Bayes' theorem assuming strong independence between the linguistic features.
- It is a simple and efficient method with reasonable accuracy.
- Usually used when the size of the training set is small.

Lexicon - Based Approach:

- Works on the assumption that the collective polarity of the sentence or a document is the sum of polarities of the individual phrases or words.
- Advantage is that labelled data is not required.

Software and Platform





Questions???





Thank You

